Indonesian Journal of Electrical Engineering and Computer Science
Vol. 40, No. 2, November 2025, pp. 978~987
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v40.i2.pp978-987

Multi-visual modality for collaborative filtering-based personalized POI recommendations

Sudarat Arthan[1], Kreangsak Tamee[1,2]
[1]Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok, Thailand
[2]Research Center for Academic Excellence in Nonlinear Analysis and Optimization, Naresuan University, Phitsanulok, Thailand

Article Info

Article history:

Received Dec 12, 2024
Revised Jul 30, 2025
Accepted Oct 14, 2025

Keywords:

Image-based recommendation
POI recommendation
Sparse data
User preferences
Visual modality

ABSTRACT

Point-of-interest (POI) recommendation systems help users discover locations that match their interests. However, these systems often suffer from data sparsity due to limited user check-in history. To address this challenge, this study proposed a novel user profiling framework that incorporates multiple visual modalities derived from user-generated photos. Three types of visual-based user profiles were constructed: image label-based, image feature-based, and a fused profile, combining both modalities through score-level fusion. We conducted extensive experiments on two real-world datasets. The results demonstrate that visual-based profiles, particularly the image feature-based profile, consistently improve recommendation performance under sparse data conditions. Although the fused profile offered stable results, it did not consistently outperform the single modality. Furthermore, performance was sensitive to the number of nearest neighbors and the amount of training data. These findings highlight the importance of modality selection and fusion strategy in visual-based POI recommendation systems.

Corresponding Author:

Kreangsak Tamee
Research Center for Academic Excellence in Nonlinear Analysis and Optimization
Naresuan University Thailand
65000 Phitsanulok, Thailand
Email: kreangsakt@nu.ac.th

1. INTRODUCTION

Point-of-Interest (POI) recommendation systems play a crucial role in helping users discover locations that align with their personal preferences and interests, particularly in an era where information about places is abundant and readily accessible [1]. However, a major challenge faced by these systems is data sparsity, which arises when user preferences are inferred primarily from behavioral data, such as check-ins. Since many users only visit and check in at a limited number of POIs, the available interaction data is often insufficient to accurately model their preferences [2]. This sparsity significantly impairs the system's ability to provide relevant and personalized recommendations.

To address this issue, recent studies have explored various forms of user-generated additional data as alternative or complementary sources of information. These include tags [3], [4], POI reviews [5], social connections [2], temporal patterns [6], and visual content from user-shared photos [7]. Among these, user-shared photos have received increasing attention due to their ability to provide rich contextual information about user preferences and POI characteristics. In particular, geo-tagged photos offer both spatial and temporal information, enabling the reconstruction of user travel histories and movement patterns [8]. The visual content embedded in these photos can be analyzed to infer user preferences, as demonstrated in [9]-[14].

Visual data derived from photos can be broadly categorized into two main modalities: image labels (e.g., tags or categories extracted via computer vision APIs) and image features (e.g., textures, shapes, or high-level descriptors learned from CNN-based models). These modalities serve complementary purposes in user profiling. Image features have been used in several studies to model latent visual preferences and improve recommendation accuracy. For example, Zhao *et al.* [9] and Wang *et al.* [10] utilized deep features from user photos to suggest personalized POIs and tours. Liu *et al.* [14] and Zhang *et al.* [15] further combined visual features with geographic influence to improve recommendation performance. In a more dynamic context, Sang *et al.* [16] leveraged visual features from geo-tagged photos together with sequential user behavior patterns, using an adaptive attention mechanism to balance visual cues with short-term and long-term check-in history. Image labels-often in the form of descriptive tags-also offer more interpretable semantic cues. Kim *et al.* [11] proposed a graph-based approach using image tags to model both general and individual user interests, which were then compared with candidate POIs for recommendation. These studies illustrate the growing potential of both image features and labels in enhancing user profiling for POI recommendation. In addition, Stefanovic and Ramanauskaite [13] also utilized object-level labels in user photos to infer preferences and connect them to suitable travel destinations. However, most existing work treats these modalities separately, and few studies have explored their integration or comparatively evaluated their effectiveness, especially in the context of collaborative filtering.

In this study, a novel approach to user profile framework was proposed that integrates multiple visual data modalities derived from user-generated photos. Specifically, we utilize three types of visual representations: image labels, image features, and a fused representation that combines both modalities. In this context, the term multi-visual modality refers not to multiple types of images, but rather to diverse forms of information extracted from the same image, aiming to enrich user profiles and better reflect individual preferences. The key contributions of this study are as follows.

- We propose a multi-visual user profiling framework that integrates image labels and image features extracted from user-generated photos.
- We conducted a comparative evaluation of three types: label-based, feature-based, and fused profile, using collaborative filtering approach, yielding key insights into the strengths and limitations of each approach for POI recommendation.
- Although the fused representation did not outperform single-modality profiles, the results offer critical insights into the interaction and potential redundancy between different visual modalities in sparse-data environments.
- Our findings highlight the importance of modality compatibility and provide practical guidelines for future multi-modal fusion strategies in POI recommendation systems.

The remainder of this paper is organized as follows. Section 2 presents the framework of the proposed approach. Section 3 describes the experimental setup. Section 4 presents the results and discusses the findings. Finally, Section 5 concludes the paper and outlines directions for future work.

2. METHOD

2.1. Overview of the proposed framework

In this study, a multi-visual user profiling framework was proposed that integrates two distinct visual modalities-image labels and image features-extracted from user-generated photos. The goal was to investigate how different representations derived from the same image can be used to model user preferences for POI recommendation, particularly under sparse data conditions. The framework consists of three main components: visual data extraction, user profile construction, and POI recommendation. A summary of the framework is illustrated in Figure 1.

2.2. Visual data extraction

User-shared photos are the primary data source for constructing visual user profiles. Each photo is processed through two distinct pipelines to extract semantic and visual representations.

2.2.1. Image label extraction

Image labels are extracted using Google Cloud Vision [17], a pre-trained computer vision service that automatically detects and assigns semantic tags to each photo. Google Cloud Vision was selected due to its proven effectiveness in providing rich, high-confidence semantic annotations across a wide range of image types [18], [19] with demonstrated robustness and scalability in prior studies involving image-based POI recommendations. The extracted tags represent high-level visual concepts such as beach, temple, or mountain, serving as interpretable representations of image content. For each photo, the top 10 labels with the highest confidence scores were selected and used to construct the user's label-based profile.

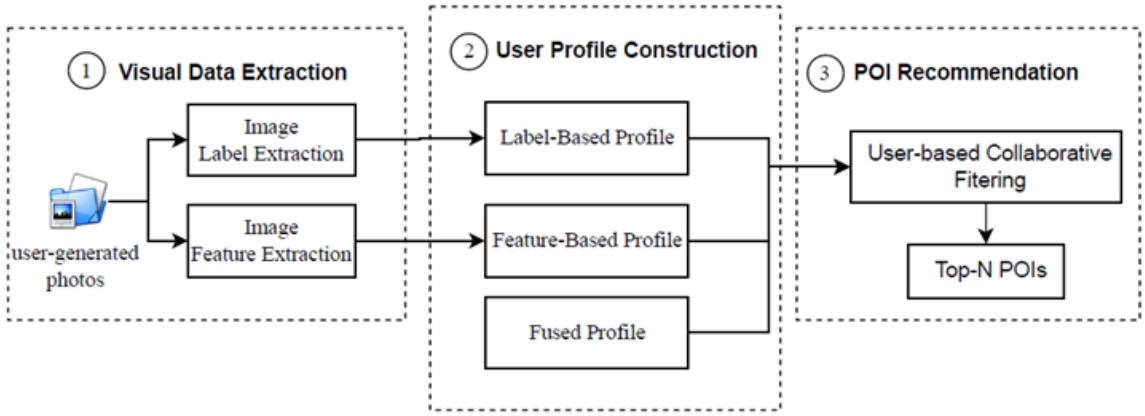


Figure 1. The overall framework of our proposed approach

2.2.2. Image feature extraction

Image features are extracted using the VGG16 convolutional neural network pre-trained on ImageNet [20]. VGG16 was chosen as the feature extractor because of its established performance in extracting meaningful visual representations for image classification tasks. Its relatively lightweight architecture and wide adoption in visual- based recommender systems [14], [16] make it a suitable choice for this study. Features are obtained from the last convolutional block of VGG16 (block5_pool) and compressed using global average pooling to reduce dimensionality and computation time, resulting in a 512-dimensional feature vector for each image.

2.3. User profile construction

We constructed three types of user profiles based on the extracted visual data: label-based profiles, feature-based profiles, and fused profiles.

2.3.1. Label-based profiles

Label-based profiles represent user interests as topic distributions derived from the detected image labels. We constructed these profiles using a topic modeling approach based on our previously proposed method [21]. The process consists of three main steps, as illustrated in Figure 2.

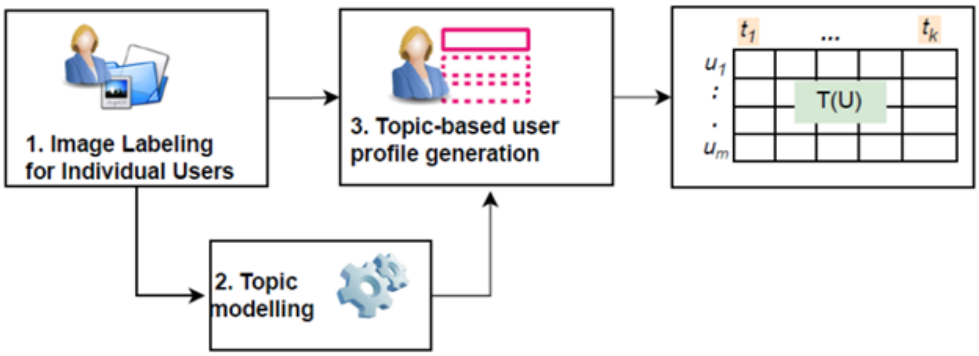


Figure 2. The process of constructing label-based user profiles

First, user-generated photos are grouped by location under the assumption that images taken at the same place reflect similar visual interests. For each user, all labels extracted from photos within each location are merged into a consolidated set, and duplicates are removed to form a unique label set per location. Second, all location-level label sets are aggregated into a corpus for topic modeling. Each location is treated as a document composed of its respective labels. We apply the latent dirichlet allocation (LDA) algorithm [22] to uncover latent topics from the corpus. To reduce noise, low-frequency labels that occur in fewer than two documents are filtered out. The optimal number of topics is selected based on a combination of quantitative

measures, such as coherence scores, and qualitative evaluation, based on tests with topic numbers ranking 2 to 20. Third, the topic distributions are computed for each visited location using the trained LDA model. A user's topic-based profile is then generated by averaging the topic distributions of all locations the user has visited. This results in a topic distribution vector representing the user's interests across latent topics.

Finally, user profiles for all users were compiled into a user-topic matrix, where each row corresponds to a user, each column to a latent topic, and each entry reflects the user's interest level in that topic.

2.3.2. Feature-based profiles

Feature-based profiles represent user interests as latent visual preferences using CNN-extracted features. The construction process consists of two main steps, as illustrated in Figure 3.

In the first step, the visual features of all photos associated with each visited location are aggregated using max pooling, as users often take multiple photos at the same place. This results in a single representative vector per location. In the second step, a user-level profile is created by applying mean pooling across all location-level vectors. The resulting vector is a 512-dimensional representation that captures the user's overall visual preferences.

Finally, user profiles for all users are compiled into a user-feature matrix, where each row corresponds to a user, each column to a visual feature dimension, and each cell contains the averaged feature value. This matrix serves as input for user similarity computation in collaborative filtering for downstream POI recommendation tasks.

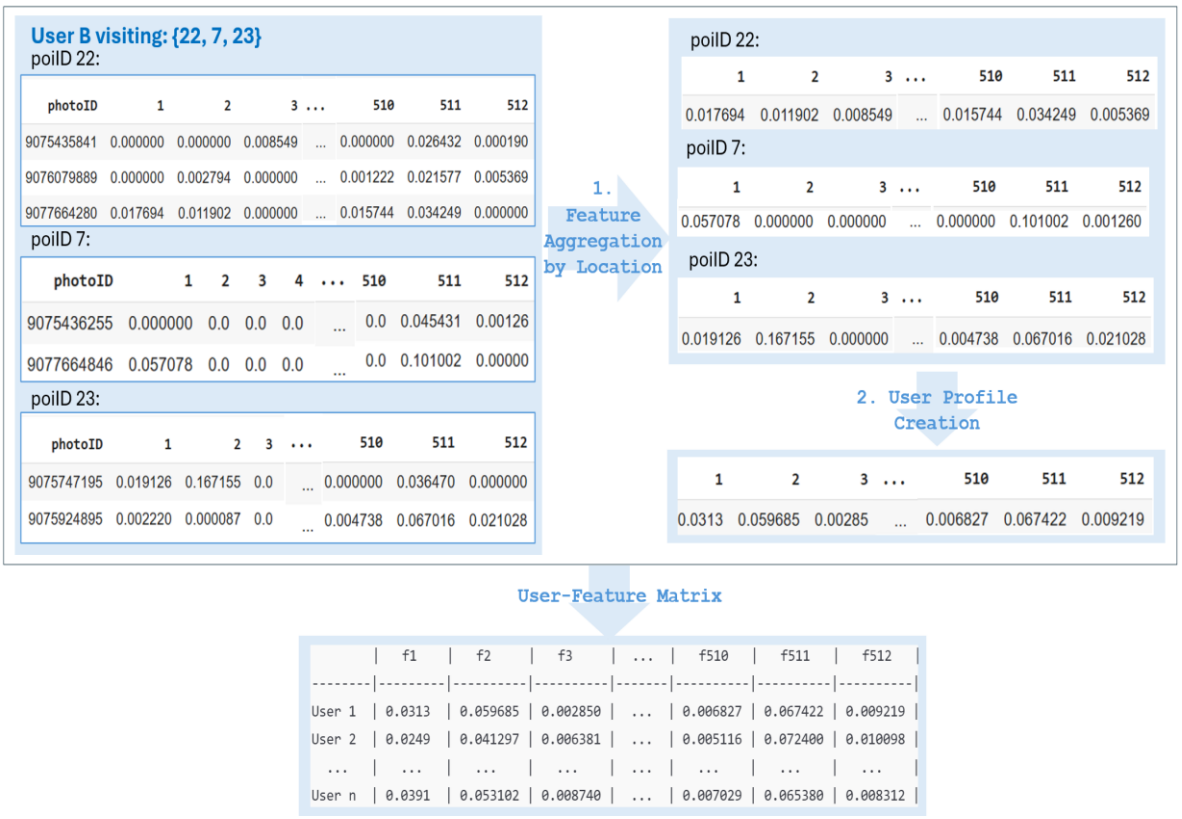


Figure 3. The process of constructing feature-based user profiles using image feature

2.3.2. Fused profiles

To capture user preferences across both semantic and visual dimensions, a late fusion strategy was adopted for combining similarities derived from label-based and feature-based user profiles. Rather than merging the raw profile vectors directly, we calculate similarity scores separately for each modality using cosine similarity [23], which is widely recognized for its robustness in high-dimensional spaces [24], [25]. Similarity scores are then integrated using element-wise multiplication. This strategy avoids the issue of combining profile vectors with inherently different scales and dimensions.

Let $U_{t,u}$ and $U_{f,u}$ denote the topic-based and visual feature-based profile vectors of user u, respectively. The cosine similarity between users u and v in each modality is computed using (1) and (2).

$$S_t(u,v) = cos(U_{t,u}, U_{t,v}) = \frac{U_{t,u} \cdot U_{t,v}}{\|U_{t,u}\| \|U_{t,v}\|} \quad (1)$$

$$S_f(u,v) = cos(U_{f,u}, U_{f,v}) = \frac{U_{f,u} \cdot U_{f,v}}{\|U_{f,u}\| \|U_{f,v}\|} \quad (2)$$

The fused similarity score is then computed by combining these modality-specific similarities using element-wise multiplication, as expressed in (3).

$$S_{fused}(u,v) = S_t(u,v) \cdot S_f(u,v) \quad (3)$$

where S_{fused} denotes the fused similarity score between users u and v, which is used for identifying the top-K most similar users to the target user in the collaborative filtering process.

2.4. POI recommendation

The final stage of the proposed framework involves generating personalized POI recommendations based on the user profiles constructed from different visual modalities. We adopted a user-based collaborative filtering approach, which identifies users with similar preferences and recommends POIs they have visited to the target user. The relevance score of a POI p for user u is estimated by aggregating the similarity-weighted interactions of the neighbors, as expressed in (4).

$$\hat{r}_{u,p} = \frac{\sum_{v \in N_k(u)} S_{fused}(u,v) * r_{v,p}}{\sum_{v \in N_k(u)} S_{fused}(u,v)} \quad (4)$$

where $\hat{r}_{u,p}$ denotes the predicted score for POI p, $r_{v,p}$ denotes a binary value indicating whether user v has visited POI p, $N_k(u)$ denotes the set of top-K similarity users to user u. The top-N POIs with the highest predicted scores are recommended to the target user.

3. EXPERIMENTAL SETUPS

3.1. Datasets

We conducted our experiments on the YFCC100M Flickr Creative Commons 100M (YFCC100M) dataset [26]. From this dataset, we selected user data corresponding to visits in Budapest and Toronto. The photos were first collected from Flickr website via the Flickr API [27]. To ensure data completeness, we excluded users whose photos could not be retrieved, as well as those who had visited fewer than five distinct POIs. The resulting dataset is summarized in Table 1.

Table 1. Summary of the filtered YFCC100M dataset used in the experiments

City	#Photos	#Users	#POIs	#POI Visits
Budapest	15,381	235	35	3,093
Toronto	30,995	209	30	3,352

3.2. Evaluation schema

To evaluate the effectiveness of the proposed multi-visual user profiling methods, we split the user data into training and testing sets. The training set sizes were varied at 20%, 40%, and 70% of each user's visited POIs, selected through random sampling of their check-in history. The remaining POIs were used as the ground truth to assess recommendation performance and examine the impact of data sparsity.

The system recommended the top-5 POIs for each user by aggregating preferences from the most similar users. To explore the impact of neighborhood size, we tested five values for the number of nearest neighbors, ranging from 30 to 50.

The quality of the recommendations was measured using two ranking-based evaluation metrics, namely Recall@5, which quantifies the proportion of relevant POIs appearing in the recommended list, and Normalized Discounted Cumulative Gain (NDCG@5), which considered both the relevance and rank position of the recommended POIs, assigning higher scores to relevant items that appear earlier in the list [16]. All metrics were computed individually for each user and then averaged across all users in the test set.

3.3. Baselines

We compared the recommendation performance across three types of visual-based user profiles: image label-based (IL), image feature-based (IF), and a fused profile referred to as ILVF2SIM, which integrates similarity scores from both modalities using element-wise multiplication. To assess the relative effectiveness of the proposed multi-visual user profiling approach, we also conducted comparative experiments with two baseline methods, summarized as follows:

- UCF: The traditional User-based Collaborative Filtering method, in which each user's profile is constructed solely based on their check-in history.
- TCM: The textual content method [4], a location recommendation framework based on User-based Collaborative Filtering method, where user profiles incorporate textual information such as POI categories and tags.

4. RESULT ANALYSIS AND DISCUSSION

4.1. Performance comparison with baseline and visual-based user profiles

In this section, we compare the recommendation performance across three types of visual-based user profiles: IL, IF, and ILVF2SIM, as well as two baseline methods. For each method and training set size, we reported the best result obtained by selecting the optimal number of nearest neighbors (NN) from the range of 30 to 50. The results based on the data in Budapest and Toronto are shown in Table 2.

Table 2. Performance comparison across user profile types

Training Size	Profile Type	Budapest				Toronto			
		Best NN	Recall@5	Best NN	NDCG@5	Best NN	Recall@5	Best NN	NDCG@5
20%	UCF	45	0.2766	40	0.4107	50	0.3174	50	0.4352
	TCM	50	0.2784	50	0.4262	50	0.3647	50	0.4859
	IL	40	0.3394	50	0.4971	50	0.3873	50	0.5084
	IF	30	0.3476	30	0.4972	50	0.3878	50	0.4999
	ILV2SIM	45	0.3376	45	0.4921	50	0.4007	50	0.5283
40%	UCF	50	0.3445	50	0.3978	50	0.4031	50	0.4339
	TCM	35	0.3368	50	0.3948	50	0.3897	45	0.4223
	IL	50	0.3832	50	0.4459	50	0.4296	50	0.4589
	IF	30	0.3821	50	0.4348	45	0.4350	45	0.4777
	ILV2SIM	50	0.3706	50	0.4333	50	0.4472	45	0.4723
70%	UCF	50	0.3781	50	0.3526	45	0.5184	45	0.4198
	TCM	40	0.3991	40	0.2948	50	0.5670	50	0.3942
	IL	30	0.3875	35	0.3365	45	0.5135	50	0.4030
	IF	50	0.3993	30	0.3299	50	0.5186	40	0.4003
	ILV2SIM	30	0.3951	30	0.3371	50	0.5289	50	0.4064

As shown in Table 2, the experimental results demonstrate that visual-based user profiling methods generally enhance POI recommendation performance, particularly under sparse training conditions such as 20% and 40% training sizes. Among the three visual profile types, the IF profile demonstrated strong performance across most settings, although it was not consistently superior to the ILVF2SIM profile in all scenarios.

In the Budapest dataset at 20% training, for example, the IF profile achieved the highest Recall and NDCG, outperforming both IL and ILVF2SIM, as well as the baseline methods UCF and TCM. In contrast, in the Toronto dataset under the same training condition, the ILVF2SIM profile achieved the highest Recall score of 0.4472, while its NDCG score of 0.4723 was slightly lower than that of the IF profile, which achieved the highest NDCG score of 0.4777. These findings highlight the benefit of leveraging visual information from user-generated photos under sparse-data scenarios.

At higher training sizes, such as 70%, however, the performance gap between visual-based profiles and the baselines narrowed. In some cases, the TCM baseline showed competitive or even superior results. For example, in the Toronto dataset at 70% training size, TCM achieved the highest Recall, surpassing all visual-based profiles. This suggests that while visual signals are particularly effective when behavioral data is limited, their advantage diminishes as richer interaction or semantic data becomes available.

These results are consistent with prior research, such as [11], [14], [16], that emphasizes the value of visual features in modeling user interests. However, unlike research that reports consistent improvements from multimodal fusion, our fused profile did not consistently outperform the single-modality profiles, such as IF or IL. This outcome may be attributed to redundancy between modalities and the limitations of our fusion strategy, which uses element-wise multiplication to combine similarity scores. Although this method helps avoid problems caused by differences in scale and structure between the two modalities, it also has several limitations, as discussed in [28], [29]. First, it assumes that both modalities are equally important at every position, which may not be true when one type of data is noisy or less informative. Second, it does not handle differences in meaning well-for example, image features usually capture visual patterns, while image labels describe high-level concepts. Finally, this method cannot capture more complex relationships between the two types of data, which limits its ability to fully combine their strengths.

4.2. Effect of training set size

Table 2 also shows that, Recall generally improved with larger training sizes. For example, the IF profile in the Toronto dataset increased from 0.3834 at 20% to 0.5186 at 70%, and the ILVF2SIM profile increased from 0.4007 to 0.5289. Similar trends were observed in the Budapest dataset. These results suggest that visual-based profiles, particularly IF, benefit from additional behavioral data and can better capture user preferences when more check-in history is available.

NDCG did not, however, follow the same increasing trend. In several cases, NDCG improved only slightly or even declined as the training data increased. For example, the ILVF2SIM profile in Toronto achieved 0.5283 at 20% training but only 0.5601 at 70%. This conflicting trend between Recall and NDCG can be attributed to the design of the data split. As the training size increases, the remaining ground truth for evaluation becomes smaller. Since NDCG is sensitive not only to the presence of relevant items but also to their positions in the ranked list, having fewer test items can lead to less informative or more volatile NDCG values. Therefore, the interpretation of the metric should consider the impact of ground truth size.

An additional observation is that the TCM baseline, while initially underperforming, showed significant improvement at higher training sizes. In Toronto, TCM achieved the highest Recall at 70% training, surpassing all visual-based profiles. However, its NDCG of 0.3942 remained slightly lower than that of all visual-based profiles, suggesting that although more relevant items were retrieved, their ranking was suboptimal.

These findings support the understanding that visual signals are highly effective under sparse-data conditions, while textual or semantic signals become more valuable as interaction data increases. Moreover, the divergence between Recall and NDCG trends underscores the importance of considering the structure of the evaluation metrics, especially the effect of training/test splits.

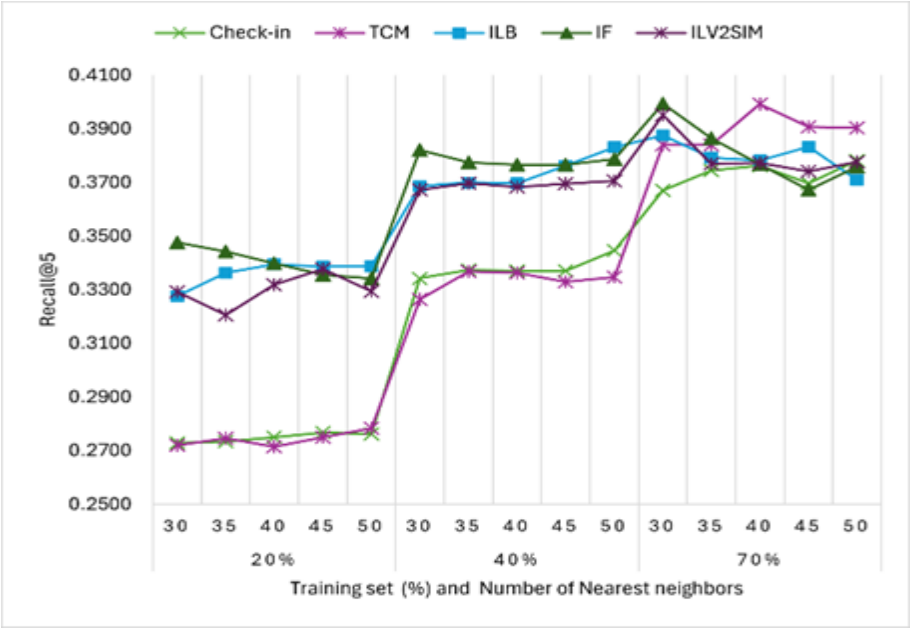
4.3. Effect of the number of nearest neighbors

To further investigate the sensitivity of performance to the number of nearest neighbors (NN), we analyzed Recall across different NN values, varying from 30 to 50 for each profile type. This purposed of this analysis was to understand how varying NN affects performance, independent of selecting the best result for comparison. The experimental results for both the Budapest and Toronto datasets are presented in Figure 4.

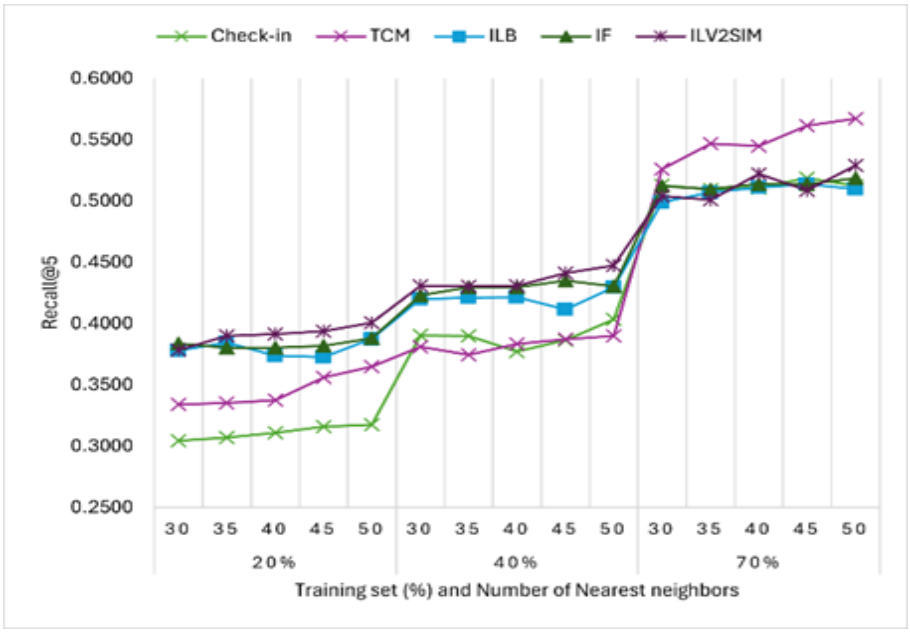
In the Budapest dataset (Figure 4a), visual-based profiles, particularly the IF profile, consistently performed best at smaller NN values, especially under sparse training conditions. For example, IF achieved the highest Recall at NN 30, while performance declined as NN increased. This trend suggests that small neighborhoods yield more relevant user similarity under data sparsity, whereas larger NN values may introduce noise from less similar users. In contrast, the Toronto dataset (Figure 4b) showed more diverse behavior. Although IF and the fused profile ILV2SIM performed competitively at smaller NN values, the TCM baseline exhibited strong gains as NN increased-reaching the highest Recall at NN 50 when the training size was 70%. This implies that text-based profiles benefit from larger user groups, particularly when more behavioral and contextual information is available to support semantic matching.

Across both datasets, the fused profile ILV2SIM demonstrated stable performance across NN values but did not consistently outperform the single-modality IF profile. This suggests that while fusion may add robustness, it does not guarantee superior accuracy unless the integration method effectively captures complementary information.

These findings indicate that the optimal number of nearest neighbors is not universal but rather profile- and context-dependent, as noted in [2], [22]. Visual-based profiles require careful tuning to avoid over-smoothing, whereas semantic-rich baselines like TCM can leverage broader neighbor sets as more training data becomes available.



(a)



(b)

Figure 4. Effect of training set size and number of nearest neighbors on recommendation performance in (a) the Budapest and (b) Toronto dataset

5. CONCLUSION

A multi-visual modality framework for POI recommendation is presented, leverge user-generated photos to construct three types of user profiles: image label-based, image feature-based, and a fused profile. The experimental results across different datasets, training set sizes, and numbers of nearest neighbors consistently demonstrated that visual-based profiles significantly enhance recommendation performance under sparse data conditions, especially the image feature-based profile. The feature-based profile showed the strongest and most stable performance, while the fused profile exhibited robustness but did not consistently outperform its single-modality counterparts. Despite the improved performance, the study also revealed limitations. The score-level fusion method, which is element-wise multiplication, assumes equal contribution from each modality and does not account for semantic differences or interaction between modalities. Additionally, the fused profiles may suffer from signal suppression when one modality is noisy or

less informative. For future work, we plan to explore adaptive or learnable fusion strategies that dynamically balance the influence of different modalities based on data quality or user context. Integrating POI descritpion or sequential user behavior may also improve personalization. Lastly, enhancing the interpretability of visual-based recommendations through explainable AI techniques can increase user trust and system transparency.

ACKNOWLEDGEMENTS

This work is partially supported by the Graduate School, Naresuan University, Phitsanulok, Thailand, through the research grant for thesis scholarships for highly potential students, Fiscal Year 2024. The authors sincerely appreciate Mr. Olalekan Israel Aiikulola, Lecturer (Special Knowledge and Abilities) at the Faculty of Medical Science, Naresuan University, Thailand, for his meticulous proofreading and editing of this manuscript. The authors also thank Mr. Roy I. Morien of the Naresuan University Graduate School for his additional editing of the grammar, syntax and general English expression in this manuscript.

FUNDING INFORMATION

Authors state no funding involved.

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] F. Ricci, "Recommender systems in tourism," in *Handbook of e-Tourism*, Cham: Springer International Publishing, 2020, pp. 1–18. doi: 10.1007/978-3-030-05324-6_26-1.

[2] K. Kesorn, W. Juraphanthong, and A. Salaiwarakul, "Personalized attraction recommendation system for tourists through check-in data," *IEEE Access*, vol. 5, pp. 26703–26721, 2017, doi: 10.1109/ACCESS.2017.2778293.

[3] B. Liu and H. Xiong, "Point-of-interest recommendation in location based social networks with topic and location awareness," in *Proceedings of the 2013 SIAM International Conference on Data Mining, SDM 2013*, Philadelphia, PA: Society for Industrial and Applied Mathematics, May 2013, pp. 396–404. doi: 10.1137/1.9781611972832.44.

[4] P. Khanthaapha, L. Pipanmaekaporn, and S. Kamonsantiroj, "Topic-based user profile model for POI recommendations," in *ACM International Conference Proceeding Series*, New York, NY, USA: ACM, Mar. 2018, pp. 143–147. doi: 10.1145/3206185.3206203.

[5] P. Mazumdar, B. K. Patra, and K. S. Babu, "Cold-start point-of-interest recommendation through crowdsourcing," *ACM Transactions on the Web*, vol. 14, no. 4, pp. 1–36, Nov. 2020, doi: 10.1145/3407182.

[6] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. Magnenat-Thalmann, "Time-aware point-of-interest recommendation," in *SIGIR 2013 - Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, Jul. 2013, pp. 363–372. doi: 10.1145/2484028.2484030.

[7] K. H. Lim, "Recommending tours and places-of-interest based on user interests from geo-tagged photos," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, New York, NY, USA: ACM, May 2015, pp. 33–38. doi: 10.1145/2744680.2744693.

[8] 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," *Knowledge and Information Systems*, vol. 54, no. 2, pp. 375-406, Feb. 2018, doi: 10.1007/s10115-017-1056-y.

[9] 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 *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, New York, NY, USA: ACM, Oct. 2017, pp. 916–924. doi: 10.1145/3123266.3123336.

[10] 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 *26th International World Wide Web Conference, WWW 2017*, Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, Apr. 2017, pp. 391-400. doi: 10.1145/3038912.3052638.

[11] K. Kim, J. Kim, M. Kim, and M. Sohn, "User interest-based recommender system for image-sharing social media," *World Wide Web*, vol. 24, no. 3, pp. 1003–1025, May 2021, doi: 10.1007/s11280-020-00832-9.

[12] F. Valls and J. Roca, "Visualizing digital traces for sustainable urban management: Mapping tourism activity on the virtual public space," *Sustainability*, vol. 13, no. 6, p. 3159, Mar. 2021, doi: 10.3390/su13063159.

[13] P. Stefanovic and S. Ramanauskaite, "Travel direction recommendation model based on photos of user social network profile," *IEEE Access*, vol. 11, pp. 28252-28262, 2023, doi: 10.1109/ACCESS.2023.3260103.

[14] B. Liu, Q. Meng, H. Zhang, K. Xu, and J. Cao, "VGMF: Visual contents and geographical influence enhanced point-of-interest recommendation in location-based social network," *Transactions on Emerging Telecommunications Technologies*, vol. 33, no. 6, Jun. 2022, doi: 10.1002/ett.3889.

[15] Z. Zhang, C. Zou, R. Ding, and Z. Chen, "VCG: Exploiting visual contents and geographical influence for Point-of-Interest recommendation," *Neurocomputing*, vol. 357, pp. 53-65, Sep. 2019, doi: 10.1016/j.neucom.2019.04.079.
[16] Y. Sang, H. Sun, C. Li, and L. Yin, "LSVP: a visual based deep neural direction learning model for point-of-interest recommendation on sparse check-in data," *Neurocomputing*, vol. 446, pp. 204-210, Jul. 2021, doi: 10.1016/j.neucom.2020.09.087.
[17] "Vision AI: image & visual AI tools," Google Cloud. Accessed: Jan. 21, 2025. [Online]. Available: https://cloud.google.com/vision
[18] A. J. Nanne, M. L. Antheunis, C. G. van der Lee, E. O. Postma, S. Wubben, and G. van Noort, "The use of computer vision to analyze brand-related user generated image content," *Journal of Interactive Marketing*, vol. 50, no. 1, pp. 156-167, May 2020, doi: 10.1016/j.intmar.2019.09.003.
[19] C. M. Rosca, A. Stancu, and M. R. Tănase, "A comparative study of azure custom vision versus google vision api integrated into ai custom models using object classification for residential waste," *Applied Sciences*, vol. 15, no. 7, p. 3869, Apr. 2025, doi: 10.3390/app15073869.
[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. arXiv:1409, p. 1556, 2015. [Online]. Available: https://arxiv.org/pdf/1409.1556
[21] S. Arthan and K. Tamee, "Leveraging user-generated visual content for point of interest recommendation," in *ICIC Express Letters, Part B: Applications*, 2024, pp. 795-802. doi: 10.24507/icicelb.15.08.795.
[22] P. Kherwa and P. Bansal, "Topic modeling: a comprehensive review," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 24, pp. 1-16, Jul. 2020, doi: 10.4108/eai.13-7-2018.159623.
[23] R. Chen, Q. Hua, Y. S. Chang, B. Wang, L. Zhang, and X. Kong, "A survey of collaborative filtering-based recommender systems: from traditional methods to hybrid methods based on social networks," *IEEE Access*, vol. 6, pp. 64301-64320, 2018, doi: 10.1109/ACCESS.2018.2877208.
[24] J. Zeng, H. Tang, Y. Zhao, M. Gao, and J. Wen, "PR-RCUC: a POI recommendation model using region-based collaborative filtering and user-based mobile context," *Mobile Networks and Applications*, vol. 26, no. 6, pp. 2434-2444, Dec. 2021, doi: 10.1007/s11036-021-01782-w.
[25] J. Zeng, F. Li, X. He, and J. Wen, "Fused collaborative filtering with user preference, geographical and social influence for point of interest recommendation," *International Journal of Web Services Research*, vol. 16, no. 4, pp. 40–52, Oct. 2019, doi: 10.4018/IJWSR.2019100103.
[26] K. H. LIM, "Datasets," Kwan Hui LIM. Accessed: Jan. 13, 2024. [Online]. Available: https://sites.google.com/site/limkwanhui/datacode
[27] F. Services, "The app garden," Flickr Services. [Online]. Available: https://www.flickr.com/services/api/
[28] Q. Zhang *et al.*, "Multimodal fusion on low-quality data: a comprehensive survey." 2024. [Online]. Available: http://arxiv.org/abs/2404.18947
[29] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–36, Oct. 2024, doi: 10.1145/3649447.

BIOGRAPHIES OF AUTHORS

Sudarat Arthan received his Ph.D. in Computer Engineering, specializing in genetic algorithms, from King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand. He is currently an Assistant Professor in the Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok, Thailand. His current interests include artificial intelligence, machine learning, optimization, and human activity recognition. He has participated in several research projects and has supervised or co-supervised more than 10 Ph.D. students. He can be contacted via email at email: sudaratar62@up.ac.th.

Kreangsak Tamee received the BA. degree in business computer from the Naresual University, Phayao, Campus, Thailand, the M.Sc. degree in Transport and Logistics management from Burapha University, Chonburi, Thailand. She is currently a Ph.D. student in information and technology, Naresuan University, Thailand. Her current research interests include recommendation systems, machine learning, and business data analytics. She can be contacted at email: kreangsakt@nu.ac.th.