MMM-GCN: Multi-Level Multi-Modal Graph Convolution Network for Video-Based Person Identification

Ziyan Liao, Dening Di, Jingsong Hao(✉), Jiang Zhang, Shulei Zhu,
and Jun Yin

Dahua Technology Co., Ltd., Hangzhou, China
hao_jingsong2022@163.com

Abstract. Video-based multi-modal person identification has attracted rising research interest recently to address the inadequacies of single-modal identification in unconstrained scenes. Most existing methods model video-level and multi-modal-level information of target video respectively, which suffer from separation of different levels and insufficient information contained in a specific video. In this paper, we introduce extra neighbor-level information for the first time to enhance the informativeness of target video. Then a Multi-Level(neighbor-level, multi-modal-level, and video-level) and Multi-Modal GCN model is proposed, to capture correlation among different levels and achieve adaptive fusion in a unified model. Experiments on iQIYI-VID-2019 dataset show that MMM-GCN significantly outperforms current state-of-the-art methods, proving its superiority and effectiveness. Besides, we point out feature fusion is heavily polluted by noisy nodes that result in a sub-optimal result. Further improvement could be explored on this basis to approach the performance upper bound of our paradigm.

Keywords: Person identification · Multi-modal · Multi biometrics · GCN · Feature fusion

1 Introduction

Person identification refers to confirming the identity of target person via biometric features, such as face recognition, person re-identification (Re-ID), speaker recognition, etc. With the development of deep learning, all these methods have achieved great success. For face recognition, ArcFace [4] reached 99.83% precision on LFW benchmark [9]. For Re-ID, Rank-1 accuracy raised to 97.1% [23]. For speaker recognition, the classification error rate is merely 0.85% [6].

Everything seems alright until trying to apply these methods to real unconstrained scenes. Face recognition is sensitive to pose, blur, occlusion, etc. Re-ID can't handle clothes changing yet. Person are not always speaking, resulting in a lack of audio features for speaker recognition. Therefore, it is necessary

D.-T. Dang-Nguyen et al. (Eds.): MMM 2023, LNCS 13833, pp. 3–15, 2023.
https://doi.org/10.1007/978-3-031-27077-2_1

to combine all these single-modal methods together to complement each other. Recently, the largest dataset for video-based multi-modal person identification task was proposed by [17]. Compared with single-modal person identification, this dataset has two additional levels of information available: (1) it contains multi-modal information to resolve the limitation of single-modal identification (multi-modal-level); (2)video with frames and richer content is used to replace still image which is commonly used in person identification (video-level). Consequently, approaches utilizing above two levels of information have been explored and achieved more meaningful results.

Current general paradigm [15,17] for video-based multi-modal person identification task models video-level and multi-modal-level information independently, then cascades them for joint training, as shown in Fig. 1. However, separate modeling for different levels leads to a restricted view, which is not conducive to capture correlations between video-level and multi-modal-level. Intuitively, it will be better to fuse information of different levels adaptively in a unified model.

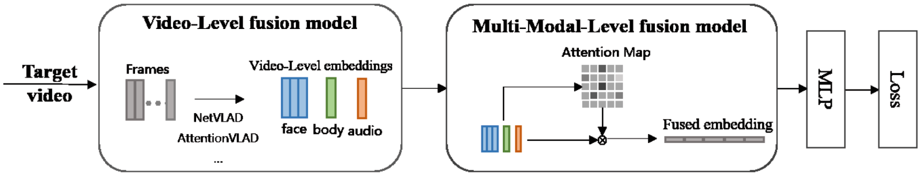

Fig. 1. Current paradigm for video-based multi-modal person identification task.

Although precision is improved by the fusion of video-level and multi-modal-level, information contained in a specific video is insufficient to determine the identity of target person, especially when visual features are broken or invisible throughout the video. Inspired by using neighbor information in re-ranking task [25], we introduce extra k-nearest neighbors for each video as its neighbor-level information. By fusing auxiliary information from neighbors, more discriminative and integral embeddings would be reconstructed.

In summary, we consider that there is some certain correlation among neighbor-level, video-level and multi-modal-level information, and hope to design an effective modeling method, which can sufficiently capture the correlation and adaptively fuse them into a unified model. Therefore, the Multi-Level Multi-Modal Graph Convolution Network (MMM-GCN) is proposed, which introduces a new paradigm for video-based multi-modal person identification task. Figure 2 shows the overall framework of MMM-GCN.

The main contributions of our work can be summarized as follows.

- For video-based multi-modal person identification task, we introduce extra neighbor-level information for the first time, to enhance the informativeness of target video.

- The proposed MMM-GCN achieves an adaptive fusion of neighbor-level, multi-modal-level and video-level in a unified model, which explores a new paradigm for video-based multi-modal person identification task.
- State-of-the-art performances are achieved by MMM-GCN on iQIYI-VID-2019 dataset [18], surpassing previous ones by a large margin. While verifying the superiority of MMM-GCN, the embedding pollution problem by noisy nodes is also pointed out, offering a clue for further research.

2 Related Work

2.1 Video-Based Multi-Modal Person Identification

Multi-modal person identification task was first proposed by [13], which also released PIPA dataset. Later, CSM dataset [10] was proposed for multi-modal person retrieval task. Due to the lack of video information or rich modalities, related approaches often use special contextual cues such as social relationships, geographical and temporal information. Therefore, the above two datasets are not suitable for video-based multi-modal person identification task. Recently, iQIYI-VID-2019 dataset [18] was released in the 2019 iQIYI Celebrity Video Identification Challenge[1]. To the best of our knowledge, it is the largest and most challenging video dataset for multi-modal person identification task. Liu et al. [17] extracted video features through NetVLAD [1] and proposed a multi-modal attention module (MMA) to fuse different modal of features adaptively. Li et al. [15] improved video-level model and multi-modal-level model respectively based on the framework of Liu et al. [17], and proposed the frame aggregation and multi-modal fusion (FAFM) method, which achieved state-of-the-art (SOTA). There are also some methods in the 2019 iQIYI Celebrity Video Identification Challenge [2,5,12]. However, most of them re-extracted features, or adopted model ensemble methods, which have less exploration on multi-modal fusion technology. In addition, all above methods modeled video-level and multi-modal-level information separately (Fig. 1), which is not conducive to exploiting correlations between different levels.

2.2 Graph Convolution Networks

Graph Convolutional Networks (GCNs) [7,11,14] extend the convolutional idea of CNNs to deal with graph data in non-Euclidean space. A graph consists of a set of objects (nodes) and relationships between these objects (edges). The basic idea of GCN is to generate a more discriminative embedding for target node by aggregating features from its neighbor nodes along edges. Due to its effectiveness and simplicity, GCNs have shown impressive capabilities on a variety of tasks [8,21,22]. Recently, some works applied GCNs to multi-modal fusion [8,20,22]. Binh et al. [20] modeled visual features and attribute labels of the body, utilized GCN to learn the topological structure of visual signature of a person. Hu et al.

[1] http://challenge.ai.iqiyi.com/detail?raceId=5c767dc41a6fa0ccf53922e7.

[8] modeled contextual information, multi-modal information and inter-speaker information in a dialogue, and used GCN to complete the task of sentiment classification of the dialogue. All these methods modeled context interaction and achieved feature aggregation through GCNs for different tasks, which verified the applicability of GCN for information fusion task.

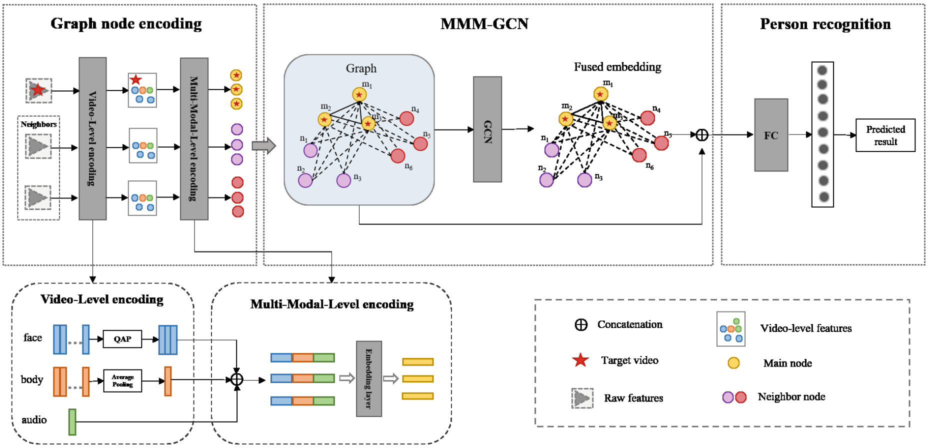

Fig. 2. Illustration of our proposed framework including three parts: (1) graph node encoding; (2) multi-level multi-modal GCN(MMM-GCN); (3) person identification.

3 Methodology

3.1 Overview

Figure 2 shows the overall framework of our method, which is sequentially divided into three parts: graph node encoding, multi-level fusion model (MMM-GCN) and person identification. Firstly, graph node encoding module(Section 3.3) encodes neighbor-level, video-level, and multi-modal-level information and generates graph node features for target video. Then, a local graph is constructed (Sect. 3.2.1) based on encoded features and adjacency containing three-level correlations, to achieve adaptive fusion of these levels through feature aggregation and update mechanism of GCN (Sect. 3.2.2). Finally, the person identification task is implemented based on the fused embedding output by GCN(Section 3.4).

3.2 Graph Construction and Learning

Different graph modeling methods have a direct effect on the information aggregation of GCNs, which further affects feature fusion. In this paper, we propose to model neighbor-level, video-level and multi-modal-level information in a unified graph, enabling target person take full advantage of the correlation among them and achieve adaptive fusion.

Graph Construction. Given a certain video, we model it with a local graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes nodes composed of target video and its neighbor videos, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is a set of edges containing correlations among video-level, neighbor-level and multi-modal-level.

Nodes. The graph is constructed as shown in Fig. 3. For each video, video-level encoding module are used to generate k_2 diverse features, which are represented by nodes with same color. Nodes with red star represent target video and the rest denote k_1 nearest neighbor videos of target video selected by neighbor-level encoding module. For each node, it obtains multi-modal information through multi-modal-level encoding module. For more details on encoding module, please refer to session 3.3. We refer to nodes of target video as main nodes, and nodes of neighbor videos as neighbor nodes, denoted by $\mathcal{M} = \{m_1, m_2, ..., m_{k_2}\}$, $\mathcal{N} = \{n_1, n_2, ..., n_{k_1 \times k_2}\}$, respectively. Thus, for a local graph, the number of graph nodes $|\mathcal{V}| = k_2 \times (k_1 + 1)$, where 1 represents target video itself.

Adjacency matrix. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ contains edge weights between each two nodes. Due to the introduction of neighbor-level information, the adjacency matrix $\mathbf{A}$ can be divided into four correlation matrices, as shown in Eq. 1

$$\mathbf{A} = \begin{bmatrix} \mathbf{MM} & \mathbf{MN} \\ \mathbf{NM} & \mathbf{NN} \end{bmatrix} \tag{1}$$

where $\mathbf{MM} \in \mathbb{R}^{k_2 \times k_2}$, $\mathbf{MN} \in \mathbb{R}^{k_2 \times (k_2 \times k_1)}$, $\mathbf{NM} \in \mathbb{R}^{(k_2 \times k_1) \times k_2}$ and $\mathbf{NN} \in \mathbb{R}^{(k_2 \times k_1) \times (k_2 \times k_1)}$ denotes the correlation matrix of $\mathcal{M} - \mathcal{M}$, $\mathcal{M} - \mathcal{N}$, $\mathcal{N} - \mathcal{M}$ and $\mathcal{N} - \mathcal{N}$, separately.

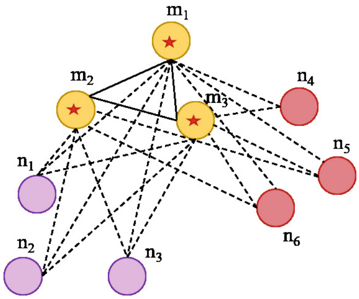

Fig. 3. Illustration of graph construction

According to the construction of graph nodes, **MM** contains both multi-modal-level and video-level correlations (solid line between nodes in Fig. 3), and **MN** adds additional neighbor-level correlations (dotted line between nodes in Fig. 3) on the basis of **MM**. When computing the elements of **MM** and **MN**, we use exponential cosine similarity and set a balance factor ω to control the fusion degree of neighbor-level information, as shown in Eq. 2, where γ denotes temperature parameter.

$$\mathbf{MM_{ij}} = exp(\frac{sim(m_i, m_j)}{\gamma}) \qquad \mathbf{MN_{ij}} = \omega\ exp(\frac{sim(m_i, n_j)}{\gamma}) \tag{2}$$

Moreover, to keep model focus on the fusion of target video, we suppress the message pass to neighbor nodes. That means all elements in **NM** and **NN** are set to 0 (no edge), except for self-connection in **NN**. In fact, we have done comparison experiments and the result shows a slight decrease in precision without suppressing the propagation of the neighbor nodes. We guess preventing neighbor nodes from aggregation could act as stabilizer and anchor to improve model stability, especially if neighbors are noisy nodes.

The adjacency matrix **A** calculated based on the above description is a non-symmetric matrix, whose degree matrix is calculated as:

$$\mathbf{D}_{row}(i,i) = \sum_j \mathbf{A}_{ij} \qquad \mathbf{D}_{col}(j,j) = \sum_i \mathbf{A}_{ij} \tag{3}$$

The normalized graph Laplacian matrix [14] is calculated as:

$$\hat{\mathbf{A}} = \mathbf{D}_{row}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}_{col}^{-\frac{1}{2}} \tag{4}$$

Graph Learning. Following chen et al. [3], we use Eq. 5 to update node embedding of the network.

$$\mathbf{F}^{l+1} = \sigma(((1-\alpha)\hat{\mathbf{A}}\mathbf{F}^l + \alpha\mathbf{F}^0)((1-\beta^l)\mathbf{I} + \beta^l\mathbf{W}^l)) \tag{5}$$

where α and β^l are hyperparameters, σ denotes the activation function, $\mathbf{W}^l$ is a learnable weight matrix. A residual connection to the first layer $\mathbf{F}^0$ is added to $\hat{\mathbf{A}}\mathbf{F}^l$, and an identity mapping $\mathbf{I}$ is added to the weight matrix $\mathbf{W}^l$. We set β^l as shown in Eq. 6, where λ is also a hyperparameter to ensure the decay of weight matrix adaptively increases when stacking more layers.

$$\beta^l = \log(\frac{\lambda}{l} + 1) \tag{6}$$

3.3 Graph Node Encoding

Neighbor-level Encoding. In this paper, we introduce extra neighbor-level information to enhance the informativeness of target video. Specifically, we construct KNN graph for target video, and encode neighbor-level information into target embeddings through message aggregating mechanism of GCN. Since face modal has the best identification effect and robustness compared to other modalities, we use face feature to calculate the k-nearest neighbors of target video based on the cosine similarity.

Video-level Encoding. Video has a large number of similar frames compared to still image. To ensure the completeness of context information while reducing redundant information and computational burden, we propose a quality-based average pooling (QAP) method to extract k_2 video-level features for each video. Specifically, due to the strong correlation between face quality and similarity, average pooling operation is performed respectively on feature groups with different quality to generate diverse video-level features. Please refer to session 4.1 for parameter settings of QAP.

Multi-Modal-Level Encoding. In order to make full use of the complementarity between multi-modal features and reduce redundancy, we designed a multi-modal-level encoding module. For a given video, we concatenate multi modal features from each video-level, then map the concatenated multi-modal feature into a unified and low-dimensional space through a fully connected layer, to get a more informative and compact feature.

3.4 Person Identification

We concatenate origin node feature n_t with the output of GCN f_t to generate the final representation for each node refer to [8]. Predicted probability is obtained through a FC layer and $softmax$ layer, as shown in Eq. 7, where $\|$ is the concatenation operation. We use categorical cross-entropy along with L2-regularization as the loss function.

$$P_t = softmax(FC(f_t \parallel n_t)) \tag{7}$$

4 Experiment

4.1 Experiment Setups

Dataset. We evaluate our proposed MMM-GCN on iQIYI-VID-2019 dataset [18]. To our best knowledge, iQIYI-VID-2019 dataset is the largest benchmark for video based multi-modal person identification. It contains 200k videos of 10k celebrities under unconstrained scenes, which makes person identification task more challenging. To ensure a fair comparison of different fusion methods, we use multi-modal features (face, head, body and audio) and face quality scores provided by official.

Implementation Details. In our best model, face, body and audio feature are used. When calculating edge weights, ω is set as 2 and γ as 1 through cross-validation. Three layers of GCN are used. The dimension of all hidden layers is set to 512. The hyperparameters α and λ are 0.1 and 0.5, respectively. We trained for a total of 40 epochs. Initial learning rate is set to 0.001, decreases by a factor of ten every 20 epochs. The Adam optimizer is used, weight decay is 0.00001.

When constructing a local graph, k_1 is set to 2. For QAP, we use a certain quality threshold 40 to classify features into high and low quality, to obtain AP_{high}, AP_{low}. The same operation is performed on all features to obtain AP_{mid}. Thus $k_2 = 3$. During testing, prediction result of AP_{high} from target video was taken as the final result. Due to small discrimination of body modal, average pooling operation is directly performed to obtain video-level features of body modal.

Evaluation Metrics. To evaluate the retrieval results, we use the Mean Average Precision (MAP) [19], which is the official metric of iQIYI-VID-2019 dataset:

$$MAP(Q) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{m_i} \sum_{j=1}^{n_i} Precision(R_{i,j}) \tag{8}$$

where Q is the set of person IDs to retrieve, m_i is the number of positive examples for the i-th ID, n_i is the number of positive examples within the top k retrieval results for the i-th ID, and $R_{i,j}$ is the set of ranked retrieval results from the top until you get j positive examples. Following the official evaluation, only top 100 retrievals are kept for each person ID.

4.2 Comparison to the State of the Art

Table 1 compares MMM-GCN to the current methods on iQIYI-VID-2019 dataset. The experimental results show that MMM-GCN significantly outperforms other methods. Compared with FAMF [15] (row 4), our method improves MAP by 2.47%, reaching 85.41% corresponds to row 6 in Table 1.

Table 1. Comparison with the state-of-the-art method on iQIYI-VID-2019 dataset.

Method	Level	MAP
GhostVLAD [24]	v+m	0.8109
Liu et,al [17]	v+m	0.8246
NeXtVLAD [16]	v+m	0.8283
FAMF [15]	v+m	0.8294
MMM-GCN	v+m	**0.8406** (+ 1.12%)
MMM-GCN	v+m+n	**0.8541** (+ 2.47%)

Table 2. MAP result of MMM-GCN under different number of neighbors.

k_1	MAP
0	0.8406
1	0.8508
2	**0.8541**
4	0.8502
9	0.8381
19	0.8281

Table 3. MAP result of MMM-GCN under different video-level features.

Video-Level Features			MAP
AP^m_{high}	AP^m_{mid}	AP^m_{low}	
✓			0.8426
	✓		0.8418
		✓	0.6859
✓	✓		0.8522
✓		✓	0.8462
✓	✓	✓	**0.8541**

Table 4. MAP result of MMM-GCN under different combinations of multi-modal.

Modal				MAP
Face	Head	Body	Audio	
✓				0.8476
	✓			0.6190
		✓		0.3949
			✓	0.2613
✓	✓			0.8448
✓		✓		0.8497
✓			✓	0.8482
✓		✓	✓	**0.8541**
✓	✓	✓	✓	0.8499

4.3 Ablation Study

Effect of MMM-GCN. All current methods (rows 1–4 in Table 1) model video-level and multi-modal-level information independently for person identification. To further verify the superiority of MMM-GCN, we experimented MMM-GCN with only video-level and multi-modal-level fusion, as shown in row 5 in Table 1. Compared with the current SOTA [15], our method improves MAP by 1.12%, which verifies the effectiveness of our proposed framework for adaptive fusion on different levels of information.

Effect of Neighbor-Level Fusion. When fusing neighbor-level information, k_1 is a hyperparameter. We hope that the selected set of neighbors contains more samples with same label, making the information richer and more comprehensive, while not introducing too much noise from neighbors with different labels that may pollute target features during fusion. The effect of neighbor selection on feature fusion will be described in detail in Sect. 4.4.

Table 2 shows effect of neighbor-level fusion, where $k_2 = 0$ means neighbor-level information is not used. It shows that $1 \leq k_2 \leq 4$ achieves a higher MAP than $k_2 = 0$, and the best result is obtained at $k_2 = 2$. When k_2 becomes larger, for example, k_2 is 9 and 19 respectively, noisy information gradually increases, which has a negative impact on MMM-GCN and reduces the fusion performance.

Effect of Video-Level Fusion. Table 3 shows the effect of video-level fusion. When using a single video-level feature, AP_{high} performs best, followed by AP_{mid} and AP_{low}, respectively. Although AP_{low} does not work well alone, it still has some unique information to improve feature diversity and raise MAP, when cooperating with other video-level features. Above results indicate that MMM-GCN could extract complementary information from video-level features adaptively, which verifies its effectiveness. In addition, MMM-GCN can also learn some valuable information between local feature, represented by AP_{high}, AP_{low}, and global feature, represented by AP_{mid}.

Effect of Multi-Modal Level Fusion. Table 4 shows the effect of multi-modal-level fusion. It can be seen that face modal performs best compared with other single modals. We take single face modal as baseline and combine it with head, body and audio respectively, results are shown in rows 5–7. Except for head modal, other modals combined with face modal could improve identification precision compared to the single face modal. We guess that the high redundancy between head and face makes the fusion invalid, while the head feature is less robust than face due to some unstable factors (hairstyle, etc.), leading to a lower MAP when fusion. Finally, we fused face, body and audio modal together, and obtained the best result with a MAP of 85.41%.

Table 5. Distribution of true and false samples under different neighbor purity (NP). Numeric in bracket represents the number of samples.

Neighbor purity	True samples	False samples
1	0.9974 (36556)	0.0026 (94)
0.5	0.7688 (2583)	0.2312 (777)
0	0.1880 (1019)	0.8120 (4402)

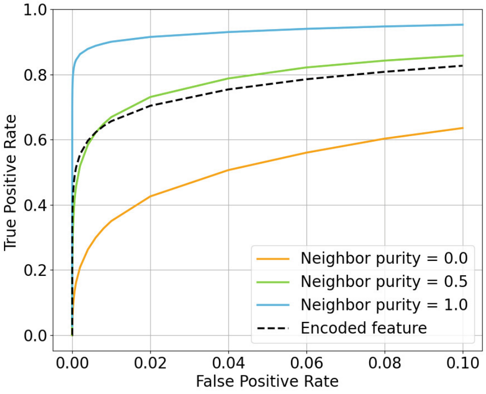

Fig. 4. ROC curves of 1:1 feature verification from randomly selected pairs, where encoded feature means the feature before fusion, the others represent fused embeddings with different neighbor purity.

4.4 Discussion

Our ablation experiments verified the effectiveness of MMM-GCN on neighbor-level, video-level and multi-modal-level fusion. However, for neighbor-level, the number of neighbors will influence performance, as shown in Table 2.

We calculated neighbor purity(NP) of true and false samples from identification results, as shown in Table 5, where NP is defined as the proportion of neighbors with same label as target video. A smaller NP means a larger number of noisy nodes (neighbors with different labels). It can be seen that with the decrease of NP, the ratio of false samples increases quickly. When $NP = 0$, false samples accounted for a majority of ratio, reaching 81.2%. The opposite performance occurs when NP increases. Above results illustrate that noisy nodes directly increase the risk of misidentification. When neighbors are selected incorrectly, the noisy information propagated by noisy nodes will confuse target features during the aggregation of GCN, thus degrading performance.

Then, we plot the ROC curves for different values of NP, as shown in Fig. 4. We use encoded feature as baseline. It is observed that performance decreases with the decrease of NP. Compared with encoded feature, $NP = 1$ shows a significant improvement, while $NP = 0$ degrades model performance.

Finally, we use official label information to manually set NP of all target videos to 1. The MAP improved by 6.86% compared with our best result, reaching 92.27%, which reconfirms that noisy nodes heavily pollute fusion, while correct neighbors can achieve an extra significant improvement.

Based on results above, we conclude that GCN is not robust enough to noisy nodes. The target feature would be heavily polluted by noisy nodes, and decreased distance between them, which makes the performance degrade. This conclusion points out a promising direction for subsequent research.

5 Conclusion

In this paper, we propose a novel framework named MMM-GCN for video-based multi-modal person identification task. Compared with existing methods, we introduce extra neighbor-level information, and achieve adaptive fusion among three levels in a unified modal. We validate the effectiveness of our proposed method on the iQIYI-VID-2019 dataset. The results show that MMM-GCN outperforms current state-of-the-art method, improving MAP by 1.12% when fusing same level information (v+m) as SOTA, and by 2.47% after introducing neighbor-level information. Finally, we point out feature fusion is heavily polluted by noisy nodes in neighbors and explore the upper bound on this direction.

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5297–5307 (2016)

2. Chen, J., Yang, L., Xu, Y., Huo, J., Shi, Y., Gao, Y.: A novel deep multi-modal feature fusion method for celebrity video identification. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2535–2538 (2019)
3. Chen, M., Wei, Z., Huang, Z., Ding, B., Li, Y.: Simple and deep graph convolutional networks. In: International Conference on Machine Learning, pp. 1725–1735. PMLR (2020)
4. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
5. Dong, C., Gu, Z., Huang, Z., Ji, W., Huo, J., Gao, Y.: DeepMEF: a deep model ensemble framework for video based multi-modal person identification. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2531–2534 (2019)
6. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S.: DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1–1.1. NASA STI/Recon technical report n 93, 27403 (1993)
7. Hamilton, W., Ying, Z., Leskovec, J.: Inductive representation learning on large graphs. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
8. Hu, J., Liu, Y., Zhao, J., Jin, Q.: MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation. arXiv preprint arXiv:2107.06779 (2021)
9. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database forstudying face recognition in unconstrained environments. In: Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition (2008)
10. Huang, Q., Liu, W., Lin, D.: Person search in videos with one portrait through visual and temporal links. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 425–441 (2018)
11. Huang, W., Zhang, T., Rong, Y., Huang, J.: Adaptive sampling towards fast graph representation learning. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
12. Huang, Z., Chang, Y., Chen, W., Shen, Q., Liao, J.: Residual dense network: a simple approach for video person identification. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2521–2525 (2019)
13. Joon Oh, S., Benenson, R., Fritz, M., Schiele, B.: Person recognition in personal photo collections. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3862–3870 (2015)
14. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
15. Li, F., Wang, W., Liu, Z., Wang, H., Yan, C., Wu, B.: Frame aggregation and multi-modal fusion framework for video-based person recognition. In: Lokoč, J., et al. (eds.) MMM 2021. LNCS, vol. 12572, pp. 75–86. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67832-6_7
16. Lin, R., Xiao, J., Fan, J.: NeXtVLAD: an efficient neural network to aggregate frame-level features for large-scale video classification. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)
17. Liu, Y., et al.: iQIYI-VID: a large dataset for multi-modal person identification. arXiv preprint arXiv:1811.07548 (2018)
18. Liu, Y., et al.: iQIYI celebrity video identification challenge. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2516–2520 (2019)
19. Manning, C., Raghavan, P., Schütze, H.: Introduction to information retrieval. Nat. Lang. Eng. **16**(1), 100–103 (2010)

20. Nguyen, B.X., Nguyen, B.D., Do, T., Tjiputra, E., Tran, Q.D., Nguyen, A.: Graph-based person signature for person re-identifications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3492–3501 (2021)
21. Shen, S., et al.: Structure-aware face clustering on a large-scale graph with 107 nodes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9085–9094 (2021)
22. Tao, Z., Wei, Y., Wang, X., He, X., Huang, X., Chua, T.S.: MGAT: multimodal graph attention network for recommendation. Inf. Process. Manag. **57**(5), 102277 (2020)
23. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 274–282 (2018)
24. Zhong, Y., Arandjelović, R., Zisserman, A.: GhostVLAD for set-based face recognition. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11362, pp. 35–50. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20890-5_3
25. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1318–1327 (2017)