Ranking Aggregation with Interactive Feedback for Collaborative Person Re-identification

Ji Huang[1,2]
2019301050286@whu.edu.cn
Chao Liang*[1,2]
cliang@whu.edu.cn
Yue Zhang[1,2]
moozy924@whu.edu.cn
Zhongyuan Wang[1,2]
wzy_hope@163.com
Chunjie Zhang[3]
cjzhang@bjtu.edu.cn

[1] National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
Wuhan 430072, China

[2] Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University
Wuhan 430072, China

[3] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University
Beijing 100044, China

Abstract

Person re-identification (re-ID) aims to retrieve the same person from a group of networking cameras. Ranking aggregation (RA), a method to aggregates multiple ranking results, can further improve the retrieval accuracy in re-ID tasks. Existing RA work can be generally divided into unsupervised methods and fully-supervised methods. Unsupervised methods lack external supervision, can hardly achieve the optimal results. In contrast, fully-supervised methods need massive labeling data for training, which is prohibitively expensive in the practical application. This paper studies interactive RA (IRA) to address the above challenges in existing RA research. The core idea is to utilize a small amount of supervisory information, obtained from users' relevance feedback, to supervise RA method to produce better re-ranking results. Compared with unsupervised methods, IRA introduces supervisory information and thus has better aggregation accuracy. Compared with fully-supervision methods, the supervisory information of IRA is more readily available, and can be targeted to specific queries. Particularly, we propose two IRA implementations, based on ranking positions and scores respectively, to adapt to diverse application scenarios, where rankers only give rankings, or rankers give similarity scores. Experiments on three public re-ID datasets have shown that IRA significantly outperforms the-state-of-art unsupervised baselines, and achieves similar accuracy with less labeling cost than the fully-supervised RA method.

*Chao Liang is the corresponding author.

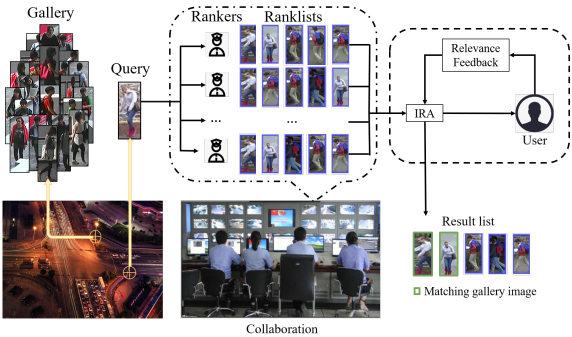


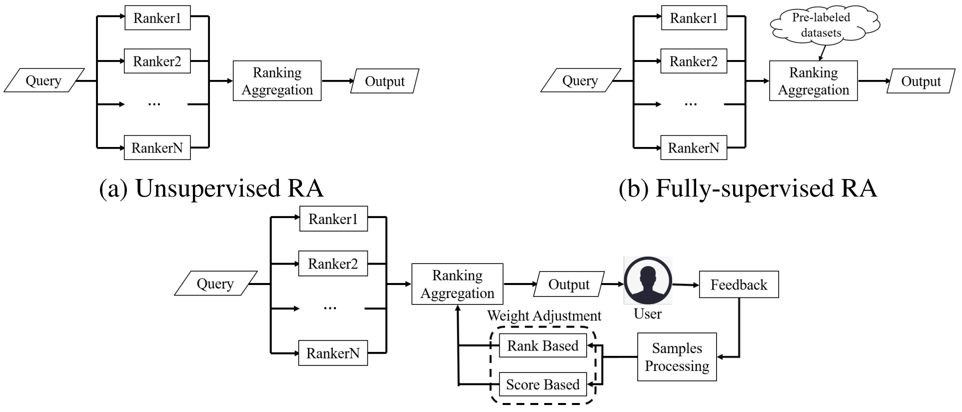
Figure 1: The multi-person collaboration re-ID.

1 Introduction

The goal of person re-identification (re-ID) is to retrieve the same person from a group of networking cameras, which has been widely used in the field of criminal investigation [14, 16, 21, 22, 25, 30]. Most previous re-ID studies focus on devising a single method to improve the retrieval accuracy, ignoring the effective fusion of different ranking methods. In the real-world scenario, video investigation missions are usually conducted by multi-person cooperation (see Fig. 1). Once a case occurs, several policemen will retrieve the video of the crime location, and identify a few highly suspected query images and a large number of gallery images. Then different policemen investigate the query from their personalized perspective. The final retrieval result are gathered by the submitted rankings of all investigators. Since different investigators have their own strengths and weaknesses, no investigator can achieve consistent superiority in all queries. In this case, it is very important to effectively aggregate the rankings submitted by different investigators. Yu *et al* [28] first notice this multi-person collaborative re-ID working mode and formulate it as a crowd sourcing based ranking aggregation (RA) problem. Inspired by their work, this paper continues to study of RA in re-ID problem, but does not require a pre-labeling training dataset that is hardly available in the practical application.

The existing research on RA includes unsupervised methods and supervised methods. Unsupervised RA [5, 6, 19, 19, 26, 26, 28], as shown in Figure 2 (a), does not use any supervisory information. These unspervised methods heavily rely on manually setting parameters for different queries. On the contrary, fully-supervised RA, as shown in Figure 2 (b), fuses multiple rankings with the guidance of the external supervision [2, 4, 28]. For example, CSRA [28] evaluates the reliability of each ranker through supervised training, and aggregates multiple rankings according to their corresponding reliability scores. But, in the real scenario, problem-oriented pre-labeled datasets are hardly available, prohibit fully-supervised RA methods from practical applications.

To address the above problems, we investigate an interactive RA (IRA) method. Its core idea is utilizing a small amount of readily available supervisory information, obtained from users' relevance feedback, to guide the dynamic adjustment of RA method to generate an improved RA result. It is a cost effective compromise, between unsupervised and fully-supervised RA methods, enjoying the advantage of excellent RA performance without intensive labeling effort. Specifically, compared with fully-supervised RA methods, IRA



(a) Unsupervised RA (b) Fully-supervised RA

(c) Interactive RA

Figure 2: Procedure of different RA methods.

doesn't need time-consuming data collection and model training beforehand. In addition, the relevance feedback is always made for a specific query, making IRA an effective and efficient therapy for the collaborative re-ID.

Our research is inspired by the seminal work of Rui et al [27]. But differently, relevance feedback is utilized to evaluate the reliability of each basic ranker for RA, rather than different features in [27]. Firstly, IRA get an initial ranking and output it to the user. Users can choose to give feedback about the top ranking results in the initial ranking. For each gallery, the user can mark positive or negative. Positive means that the gallery image sharing the same identity as the query, while negative means different. Then, IRA uses the feedback supervisory information to measure the reliability of each single ranker and weights them. After, all rankers are aggregated again according to their weights. In addition, IRA will place positive samples at the top and negative samples at the bottom, to obtain the new fusion ranking result. Above process can be repeated until user no longer interacts. As far as we know, IRA is the first interactive RA work for re-ID problem. For the step of adjusting weights in IRA, we propose two implementations, ranking based IRA and score based IRA, with a unified framework, to adapt to two different interaction scenarios where rankers only give rankings without similarity score [7], or rankers give similarity scores [19].

In general, the main contributions of this paper are summarized as follows:

- We propose an IRA method for collaborative re-ID problem. Compared with unsupervised and fully-supervised RA methods, IRA enjoys the advantage of excellent RA performance without intensive labeling effort.

- We designe two IRA implementations, ranking-based IRA and score-based IRA, to adapt to diverse ranking scenarios with or without ranking scores.

- We compare IRA with both unsupervised RA methods and fully-supervised RA method on CUHK03 [12], Market1501 [29], DukeMTMC-reID [20] datasets, to validate both effectiveness and efficiency of our proposed method.

2 Proposed Method

2.1 Problem statement

Given a query image q and J galleries $\left\{g_j\right\}_{j=1}^{J}$, I rankers $\{r_i\}_{i=1}^{I}$ retrieve q respectively as $s_{ij} = R_i(q, g_j)$, where $R_i()$ is the retrieve function of ranker r_i that returns the similarity score between q and g_j. The $s_i \in \mathbb{R}^J$ is the normalized similarity score vector of $\left\{g_j\right\}_{j=1}^{J}$ returned by ranker r_i, and $\mathbf{S} = \{s_i\}_{i=1}^{I} \in \mathbb{R}^{I\times J}$ is the similarity score matrix of all rankers. With $\mathbf{S}$, ranklist $\mathbf{L} = \{l_i\}_{i=1}^{I} \in \mathbb{R}^{I\times J}$ can be obtained as: $l_i = sort(s_i)$, in which $l_i \in \mathbb{R}^J$ is the ranklist given by r_i.

All the galleries with positive user feedback will be marked as $\left\{g_k^+\right\}_{k=1}^{K}$, and negative galleries will be marked as $\left\{g_p^-\right\}_{p=1}^{P}$. The positive and negative feedback samples of interaction round m are expressed as g_m^+ and g_m^-. Total interaction rounds is M, $\mathbb{G}_M^+ = g_1^+ \cup g_2^+ \cup \ldots \cup g_M^+$ and $\mathbb{G}_M^- = g_1^- \cup g_2^- \cup \ldots \cup g_M^-$ are the sets of all positive and negative samples respectively. The goal of our method is to aggregate the similarity score matrix $\mathbf{S}$ into a final score vector r_M^* based on relevance feedback $\mathbb{G}_M^+$ and $\mathbb{G}_M^-$.

2.2 Get initial ranking

Before starting user feedback, we first need to get an initial ranking and show it to users. At best, users can get satisfactory result without feedback, so the initial ranking should have high accuracy. We choose the Mean method [1] to get the initial ranking: $r_0 = \frac{1}{I}\sum_{i=1}^{I} s_i$. Using the Mean method to get the initial ranking has a small cost. In the experiment, we observed that Mean has good performance.

2.3 Relevance feedback

With the initial RA result s_0^*, galleries are displayed to user in order of similarity from high to low. User can mark the top K galleries with the highest similarity. For each gallery, the user can choose to mark postive or negative, postive means that the person in the gallery and the query images are the same one, and negative means not. Positive galleries will be placed at the top of the ranklist and negative galleries at the bottom, because the positive galleries given by user is most likely to be the groundtruth.

Note that the galleries marked by the user in each round are not repeated. The system will find the K top unmarked galleries for interacting.

2.4 Weight adjustment

After t rounds of interaction, the weights of rankers is $w^M = \left\{w_i^M\right\}_{i=1}^{I}$. Inspired by Rui's work [27], we propose two implementations to adjust the weight: ranking based IRA and score based IRA. Both implementations adjust the weight based on the positive samples. When no positive samples are fed back, the weight will remain unchanged.

The ranking based IRA adjusts the weight by positiosn of the positive samples in the ranking given by each ranker. The weight is calculated as follow:

$$w_i^M = \sum_{g\in\mathbb{G}_M^+} f(\varphi(g, l_i)) \tag{1}$$

	Aligned	Cam	Densenet	GOG	HACNN	HHL
Euclidean	52.32	25.85	37.28	0.79	43.03	27.90
Kissme	50.77	26.18	36.43	11.07	43.35	26.05
kLFDA	43.08	22.82	28.67	16.50	33.61	21.81
LMNN	44.03	13.62	32.27	2.47	33.77	18.33
Mahalanobis	50.76	23.98	36.47	4.75	43.00	26.63
NFST	45.76	26.09	28.66	17.51	34.83	28.98
OASIS	50.86	25.17	33.29	5.10	41.18	5.19
XQDA	53.11	29.01	36.50	1.54	44.30	13.72

Table 1: Re-ID mAP performance (%) of 48 basic rankers on CUHK03LABELED.

in which $\varphi(g,l_i)$ computes the ranking of g in l_i. The $f(\varphi(g,s_i))$ is a is a function that converts the ranking position to score. Placing positive galleries at a higher position means the ranker is more reliable, so the higher g is ranked, the higher ranker's score $f(\varphi(g,l_i))$ is.

Specially, Yu's work[28] has mentioned the importance of top ranking in RA problem, so f is designed as follow: $f(\varphi(g,l_i)) = \frac{1}{\lceil \varphi(g,l_i)/n \rceil}$, where n is the number of galleries fed back by the user in each round. Because the number of galleries fed back by users indicates the number of galleries users are willing to browse, the highest score will be obtained if the groundtruth is ranked in top n position, and then decrease in turn.

The score based IRA is to calculate the standard deviation of the scores of all positive galleries, from which fusion weights can be computed as:

$$w_i^M = \frac{1}{\sqrt{\frac{1}{K}\sum_{g\in\mathbb{G}_M^+}[(\phi(g,s_i) - \frac{1}{K}\sum_{g'\in\mathbb{G}_M^+}\phi(g',s_i))^2] + \beta}} \tag{2}$$

where $\phi(g,s_i)$ is the similarity score of g in s_i, and the denominator is the standard deviation of the similarity score given by r_i to $g \in \mathbb{G}_M^+$. The w_i^M is in inverse proportion of the standard deviation of all similarity scores in s_i'. Small standard deviation indicates that the ranker can extract the feature of the positive galleries. A lower standard deviation represents a higher reliability, thus a higher weight. Note that the algorithm will only update the weight when there are more than one positive galleries, and in order to avoid the denominator being 0, a very small number $\beta = 0.01$ is added to the denominator.

2.5 Weighting aggregation

The final score vector s^* is the sum of each ranker's score vector multiplied by its weight as $r_M = \frac{w^M \times \mathrm{S}}{I}$, where w^M is the weight vector after M rounds of interaction, S is the initial similarity score matrix and I is the number of rankers. The $r_M \in \mathbb{R}^l$ is the final similarity score vector of this round, and will be shown to the user after sorting.

The method to get the initial ranking introduced in Section 2.2 can be regarded as equal weight aggregation, so the initial weight of each ranker is 1.

3 Experiment

In this section, we evaluate IRA performance, with intensive comparative experiments on three popular public re-ID datasets: Specifically, we first introduce the datasets and criteria

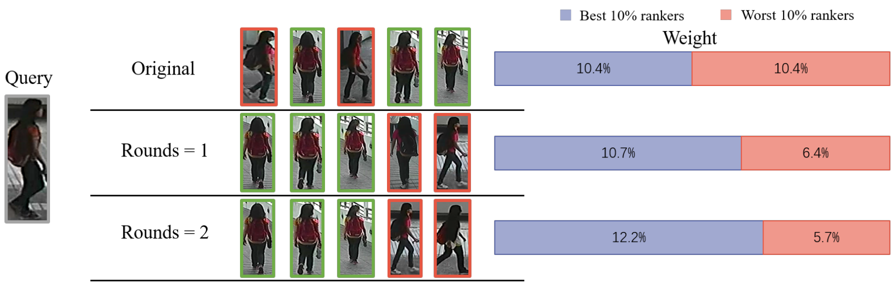


Figure 3: An retrieval example of $\mathrm{IRA}_R(3,n)$ on CUHK03LABELED, pedestrians in the green box is the groundtruth of the query, and the red is not.

used in Section 3.1, and then elaborate the preparation of basic ranker models and their original performance in Section 3.2. Comparative results with unsupervised and fully-supervised RA methods are reported in Section 3.4 and Section 3.5, respectively.

3.1 Datasets and Criteria

We evaluated IRA on three popular re-ID datasets: CUHK03 [12], Market1501 [29], DukeMTMC-reID [20]. There are two forms of CUHK03 dataset: CUHK03LABELED and CUHK03DETECTE The pedestrian frame of CUHK03LABELED is manually marked, and the pedestrian frame of CUHK03DETECTED is given by pedestrian detector. For all three datasets, we use the training set in official protocol to train ranker introduced in Section 3.2.

The datasets partition of our experiment is shown in Figure 4. For all three datasets, we use the officially divided training set to train rankers, half of the query and all galleries in the test set to train fully-supervised RA method, and the other half of the query and all galleries for testing. We use mAP and CMC@1 as indicators to evaluate ranking.

3.2 Basic rankers

To simulate single ranking models in re-ID problem, we selected 6 popular feature extraction methods (AlignedReid [24], CamStyle [31], Densenet-121 [9], GOG [18], HACNN [11], HHL [32]) and 8 popular metrics (Euclidean, Kissme [10], kLFDA [8], LMNN [23], Mahalanobis [17], NFST [13], OASIS [3], XQDA [15]). Each pair of feature and metric is regarded as a ranker, and can independently query the image. For the features and metric that need to be trained, we use the officially divided training set in the dataset for training.

In order to evaluate the performance of these features and metrics, we combined each features with each metrics, retrieve on CUHK03LABELED to evaluate the accuracy of ranking. The results are shown in the Table 1. It can be seen that there are significant performance discrepancy among different features and metrics. Hence, successful fusion of these diverse basic ranker models is not a trivial problem for collaborative re-ID problem.

3.3 Experiment Setup of IRA

For our method, we try several different feedback strategies: set the number of each round of feedback to 1, 3 and 5, because the number of annotations that users can accept is different

DATASET		MARKET1501		DukeMTMC		CUHK03 DETECTED		CUHK03 LABELED	
METHOD	SOURCE	CMC@1	mAP	CMC@1	mAP	CMC@1	mAP	CMC@1	mAP
MEAN	JMLR'11	93.82	84.46	85.01	71.00	48.71	44.54	52.71	49.73
MEDIAN	SIGMOD'04	93.94	84.06	84.74	70.21	46.71	43.20	52.71	48.77
HPA	ECIR'20	92.99	79.79	82.68	68.20	56.00	52.09	57.43	52.97
PNDCG	ECIR'20	92.16	76.89	81.51	62.46	51.71	46.99	49.71	44.70
ER	OMEGA'20	85.15	60.83	68.94	51.63	20.29	17.68	38.57	36.07
$\text{IRA}_R(1,1)$		93.76	84.31	85.64	70.89	50.43	46.60	53.00	50.46
$\text{IRA}_R(3,1)$		96.02	85.25	87.97	72.24	53.00	48.58	55.57	51.66
$\text{IRA}_R(5,1)$		**97.57**	**86.15**	**90.31**	**73.27**	**61.29**	**53.15**	**63.86**	**55.73**
$\text{IRA}_S(1,1)$		93.82	84.46	85.01	71.00	48.71	44.54	52.71	49.73
$\text{IRA}_S(3,1)$		96.14	84.76	87.16	70.79	50.71	45.98	54.86	50.95
$\text{IRA}_S(5,1)$		**97.62**	**85.76**	**89.14**	**72.44**	**60.00**	**51.23**	**63.43**	**55.67**

Table 2: Comparison of CMC@1 and mAP (%) betwenn IRA and unsupervised method. Red represents the highest value, and blue represents the second highest value.

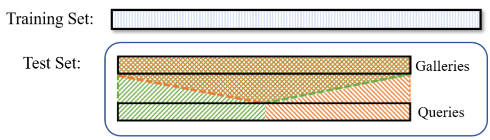


Figure 4: Dataset partition for the comparative experiment between IRA and CSRA.

in different application scenarios. For each feedback strategy, we conducted five rounds of interaction. We test two implementations of IRA. IRA_R denotes ranking based IRA, and IRA_S score based IRA. The amount of feedback items and different interaction rounds are represented by m and n. For example, $\text{IRA}_R(m,n)$ means labeler feedback m samples per round, and totally conducts n rounds of interaction.

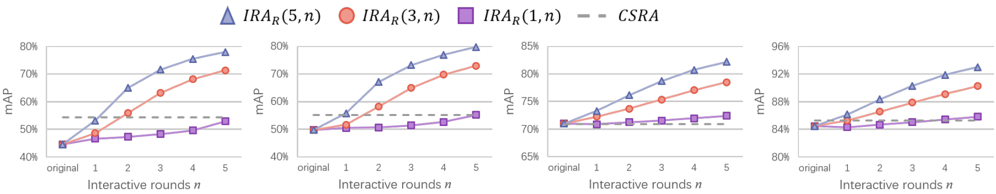
Figure 3 shows an query example of how the aggregated ranking changes with user feedback. In order to verify the effectiveness of weight adjustment, we recorded the change of weights in the retrieval process, shown in the Figure 3. It can be seen that with the increase of feedback rounds, positive galleries are gradually ranked to the highest position. At the same time, the weights of best 10% rankers remain high, while the weights worst 10% rankers continue to decrease. IRA effectively distinguishes the performance differences between rankers and gives them corresponding weights to get optimal aggregation results.

3.4 Comparison with unsupervised RA methods

Baselines: We compare IRA with averaging methods Mean [1], Median [6], weighting method ER [19], HPA [7], selection method PostNDCG [7].

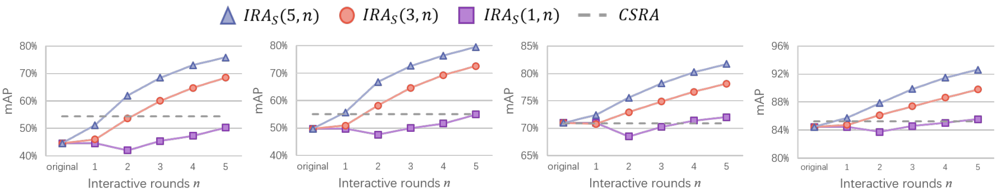
Comparison: We compared the results of IRA after one round of feedback with unsupervised methods, Because it is compared with unsupervised methods, and IRA has good accuracy after only one round of feedback in most cases. The results are shown in Table 2. For $\text{IRA}_S(1,1)$, there is only one feedback sample and the standard deviation cannot be calculated, so the result is same as Mean. It can be observed that in the case $m=3$ or 5, both IRA_R and IRA_S achieve very good accuracy, which exceeds all unsupervised RA methods.

It is observed that although the process of Mean method is very simple, it has high



(a) CUHK03DETECTED (b) CUHK03LABELED (c) DukeMTMC (d) MARKET1501

Figure 5: Results of IRA_R and CSRA on all datasets.



(a) CUHK03DETECTED (b) CUHK03LABELED (c) DukeMTMC (d) MARKET1501

Figure 6: Results of IRA_S and CSRA on all datasets.

performance on our problem, which exceeds many methods that need higher costs. Using Mean to obtain the initial ranking of IRA reduces the number of interaction rounds required.

3.5 Comparison with fully-supervised RA method

Baselines: We choose an effective fully-supervised RA method CSRA [28] to compare with our method, because CSRA is also a RA method for re-ID problem. CSRA is a fully-supervised RA method that needs training to obtain the reliability of rankers. Following the original protocol of [28], we used all galleries and half of queries in the test set for CSRA training, and compare IRA and CSRA on the other half of the test set, as shown in Figure 4.
Performance comparison: As mentioned above, we set $m = 1$, 3 and 5. IRA_R and IRA_S carry out five rounds of feedback respectively. The results of IRA and CSRA are shown in Figure 5 and Figure 6. In the case $m = 5$, IRA achieves mAP similar to CSRA with only one round of feedback. For $m = 3$, IRA can achieve higher accuracy than CSRA with two rounds of feedback. When $m = 1$, about five rounds of feedback are required to achieve an accuracy similar to that of CSRA, because the number of interactions each rounds is too small. We also note that IRA can bring some improvement on larger datasets (MARKET1501 and DukeMTMC), while CSRA is very small.

In the case of feedback one gallery per round, we observe that the performance of IRA_S may decline with feedback. The reason is that there are gallery pictures taken by the same camera as query in the ranking. This junk images' similarity score is much higher than other images, but is not included in the final measurement of performance. However, when the user marks less than 3 images, it is likely to be the junk images, which introduces noise for IRA_S, resulting in performance degradation.
Time comparison: We compared the time of IRA and CSRA on CUHK03LABELED. For

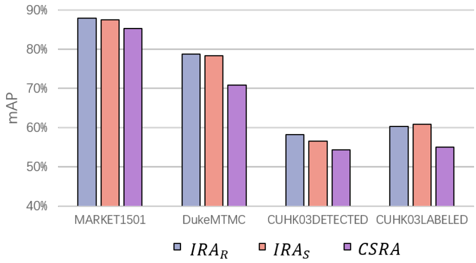


Figure 7: The mAP of IRA and CSRA with same labeling cost.

Method	mAP(%)	Train/Interact(s)	Aggregate(s)	Average(s)
CSRA	55.09	16.47	0.22	16.69
$\text{IRA}_R(1,5)$	55.21	15.00	0.03	15.03
$\text{IRA}_R(3,2)$	58.18	18.00	0.03	18.03
$\text{IRA}_R(5,1)$	55.73	15.00	0.01	**15.01**
$\text{IRA}_S(1,5)$	54.99	15.00	0.02	**15.02**
$\text{IRA}_S(3,2)$	58.22	18.00	0.02	18.02
$\text{IRA}_R(5,1)$	55.67	15.00	0.01	**15.01**

Table 3: Average time cost of IRA and CSRA to reach similar mAP on CUHK03LABELED, where red represents the shortest time cost, and blue represents the second shortest.

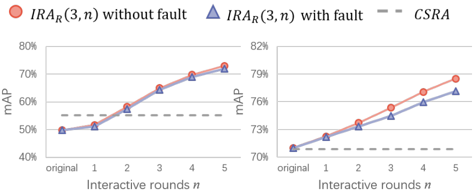
CSRA, time cost is divided into two parts: the time spent on training and time spent on aggregating. The time cost of IRA is also divided into two parts: the time spent on interacting and the time of aggregating. In our experiment, the average time for users to feedback a gallery is about 3 seconds. We estimate the average interaction time according to the m and n, and add it to IRA aggregate time to get the total time cost. For $m = 1$, 3 and 5, the change of mAP is discrete with the increase of interaction rounds. Therefore, we compared the time cost by IRA and CSRA to achieve similar mAP, which is shown in Table 3.

For CSRA, most of the time is spent on training with pre-labeled datasets, and IRA spends most of time on interacting, depending on the amount of interactions needed. Fewer interactions can significantly reduce the time spent. Compared with CSRA, IRA is more suitable for the situation where a small number of queries need to be aggregated, which often happens in practical scenario. The fusion can be started immediately without training, the results can be obtained faster.

Cost comparison: We also compared the labeling cost of IRA and CSRA, taking the average number of labels required for each test query as the measurement standard. For CSRA, we take the number of galleries in the training set as the total labeling cost. The average number of labels of CSRA is shown in Table 4.

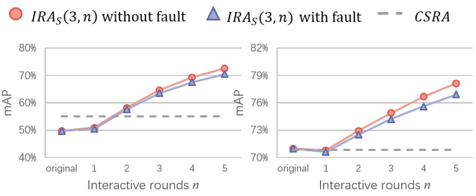
We also compared the performance of IRA and CSRA under the same number of labels. We set m to the same number as average labeling cost of CSRA, and conduct one round of interaction. For the fairness of comparison, the feedback quantity of IRA on the MARKET1501 is set to 9, is set to 15 on DukeMTMC, and is set to 7 on CUHK03DETECTED and CUHK03LABELED. The results are shown in Figure 7. Using the same amount of supervision, IRA_R and IRA_S achieve better mAP than CSRA on all datasets, because the supervision of IRA is more targeted for specific query.

In the above experiments, we simulated user feedback. Because for big datasets like



(a) CUHK03LABELED (b) DukeMTMC

Figure 8: Results of IRA_R with fault.



(a) CUHK03LABELED (b) DukeMTMC

Figure 9: Results of IRA_S with fault.

Dataset	MARKET1501	DukeMTMC	CUHK03 DETECTED	CUHK03 LABELED
Train	15,913	17,661	5,332	5,328
Test	1,684	1,114	700	700
Average	9.45	15.85	7.62	7.61

Table 4: Average labeled samples for CSRA

DATASET	DukeMTMC		CUHK03LABELED	
METHOD	CMC@1	mAP	CMC@1	mAP
$\text{IRA}_R(3,3)$	93.54%	75.35%	75.00%	64.91%
$\text{IRA}_R(3,4)$	**94.79%**	**77.06%**	**80.14%**	**69.74%**
$\text{IRA}_R(5,2)$	94.17%	76.13%	77.14%	67.05%
$\text{IRA}_S(3,3)$	93.63%	74.89%	73.86%	64.63%
$\text{IRA}_S(3,4)$	**94.52%**	**76.66%**	**78.29%**	**69.24%**
$\text{IRA}_S(5,2)$	94.17%	75.59%	76.14%	66.76%

Table 5: Accuracy of $\text{IRA}(3,3)$, $\text{IRA}(3,4)$ and $\text{IRA}(5,2)$.

DukeMTMC and MARKET1501, it's very difficult to have real users giving a lot of feedback. However, users may make error feedback in real interactions. So we tested the impact of error feedback on IRA on CUHK03LABELED and DukeMTMC, let each feedback give error answers with a probability of 2%, and the result are shown in Figure 8 and Figure 9. A small amount of error feedback slightly reduces the retrieval accuracy of both IRA_R and IRA_S. But after two rounds of interaction, IRA still achieved higher accuracy than CSRA.

We also recorded the results of IRA after more feedback, as shown in Table 5. Under different strategies, $\text{IRA}(3,4)$ gets the best results, which is mainly because it has most interactions.

4 Conclusion and future work

We study an interactive IRA method to address collaborative re-ID prlblem, and propose two implementations of IRA using ranking positions and scores respectively. Extensive experiments show the superiority of IRA over both unsupervised and supervised RA methods in ranking accuracy, operation time and labeling cost.

In the future, we will process relevance feedback in a fully data-driven way, and apply the IRA method to more practical problems besides collaborative re-ID.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. U1903214, 61876135, 62072026), the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) (GML-KF-22-16), Guangdong-Macau Joint Laboratory for Advanced and Intelligent Computing (2020B121203000 and Chinese Association for Artificial Intelligence (CAAI)-Huawei MindSpore Open Fund. we gratefully acknowledge the support of MindSpore, CANN (Compute Architecture for Neural Networks) and the Ascend AI Processor used for this research, and codes using MindSpore will also be released at https://gitee.com/chunjie-zhang/BMVC2022. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Cjc Burges, K. M. Svore, P. N. Bennett, A. Pastusiak, and W. Qiang. Learning to rank using an ensemble of lambda-gradient models. *Journal of Machine Learning Research*, 14:25–35, 2011.

[2] I. Caragiannis, X. Chatzigeorgiou, G. A. Krimpas, and A. A. Voudouris. Optimizing positional scoring rules for rank aggregation. *Artificial Intelligence*, 267, 2016.

[3] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *The Journal of Machine Learning Research*, 2010.

[4] K. Y. Chiang, C. J. Hsieh, and I. Dhillon. Rank aggregation and prediction with item features. 2017.

[5] I. C. Dourado, Daniel Carlos Guimaraes Pedronette, and Rds Torres. Unsupervised graph-based rank aggregation for improved retrieval. *Information Processing & Management*, 56(4):1260–1279, 2019.

[6] R. Fagin, S. Ravikumar, and D. Sivakumar. Efficient similarity search and classification via rank aggregation, 2004.

[7] Soichiro Fujita, Hayato Kobayashi, and Manabu Okumura. Unsupervised ensemble of ranking models for news comments using pseudo answers. 2020.

[8] M. Gou, X. Fei, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. *Springer, Cham*, 2014.

[9] G. Huang, Z. Liu, Vdm Laurens, and K. Q. Weinberger. Densely connected convolutional networks. In *IEEE Computer Society*, 2016.

[10] M Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[11] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[12] Wei Li, Rui Zhao, Tong Xiao, and X. G. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Computer Vision & Pattern Recognition*, 2014.

[13] Z. Li, X. Tao, and S. Gong. Learning a discriminative null space for person re-identification. *IEEE*, 2016.

[14] C. Liang, B. Huang, R. Hu, C. Zhang, X. Jing, and J. Xiao. A unsupervised person re-identification method using model based representation and ranking. *ACM MM*, 2015.

[15] S. Liao, H. Yang, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 2017.

[17] P. C. Mahalanobis. On the generalized distance in statistics. *proceedings of the national institute of sciences*, 1936.

[18] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] M. Mohammadi and J. Rezaei. Ensemble ranking: Aggregation of rankings produced by different multi-criteria decision-making methods. *Omega*, page 102254, 2020.

[20] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. *European Conference on Computer Vision*, 2016.

[21] Weijian Ruan, Jun Chen, Yi Wu, Jinqiao Wang, Chao Liang, Ruimin Hu, and Junjun Jiang. Multi-correlation filters with triangle-structure constraints for object tracking. *IEEE Transactions on Multimedia*, 21(5):1122–1134, 2019.

[22] Weijian Ruan, Wu Liu, Qian Bao, Jun Chen, and Tao Mei. Poinet: Pose-guided ovonic insight network for multi-person pose tracking. In *the 27th ACM International Conference*, 2019.

[23] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(1):207–244, 2009.

[24] Z. Xuan, L. Hao, F. Xing, W. Xiang, and S. Jian. Alignedreid: Surpassing human-level performance in person re-identification. 2017.

[25] M. Ye, J. Chen, Q. Leng, L. Chao, and K. Sun. Coupled-view based ranking optimization for person re-identification. In *International Conference on Multimedia Modeling*, 2015.

[26] M. Ye, L. Chao, Y. Yi, W. Zheng, Q. Leng, C. Xiao, J. Chen, and R. Hu. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia*, 18(12):2553–2566, 2016.

[27] R. Yong, T. S. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in mars. In *Proceedings of International Conference on Image Processing*, 1997.

[28] Yinxue Yu, Chao Liang, Weijian Ruan, and Longxiang Jiang. Crowdsourcing-based ranking aggregation for person re-identification. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[29] L. Zheng, L. Shen, T. Lu, S. Wang, and T. Qi. Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[30] W. Zheng, R. Hu, Y. Yi, L. Chao, and W. Huang. Multi-level fusion for person re-identification with incomplete marks. In *Acm International Conference on Multimedia*, 2015.

[31] Z. Zhong, Z. Liang, Z. Zheng, S. Li, and Y. Yi. Camera style adaptation for person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[32] Z. Zhong, L. Zheng, S. Li, and Y. Yang. Generalizing a person retrieval model hetero- and homogeneously. *European Conference on Computer Vision*, 2018.