WHU-NERCMS @ TRECVID 2021: INSTANCE SEARCH TASK

Chao Liang
cliang@whu.edu.cn

National Engineering Center for Multimedia Software
School of Computer Science, Wuhan University

December 10, 2021

Outline

- Introduction
- Approach
- Results
- Conclusion

Outline

- Introduction
- Approach
- Results
- Conclusion

Introduction

- Person-Action Instance Search (P-A INS)
  - Specific person doing specific action
  - 4 automatic runs and 4 interactive runs

Topic ID	Person	Action
9319 9320 9321	Max Stacey Peggy	sit on couch
9322 9323 9324	Stacey Bradley Shirley	holding glass
9325 9326 9327	Bradley Shirley Peggy	holding phone
9328 9329	Max Peggy	carrying bag

Topic ID	Person	Action
9330 9331 9332	Pat Shirley Peggy	holding paper
9333 9334	Max Stacey	kissing
9335 9336	Bradley Pat	open door enter
9337 9338	Max Stacey	holding cloth

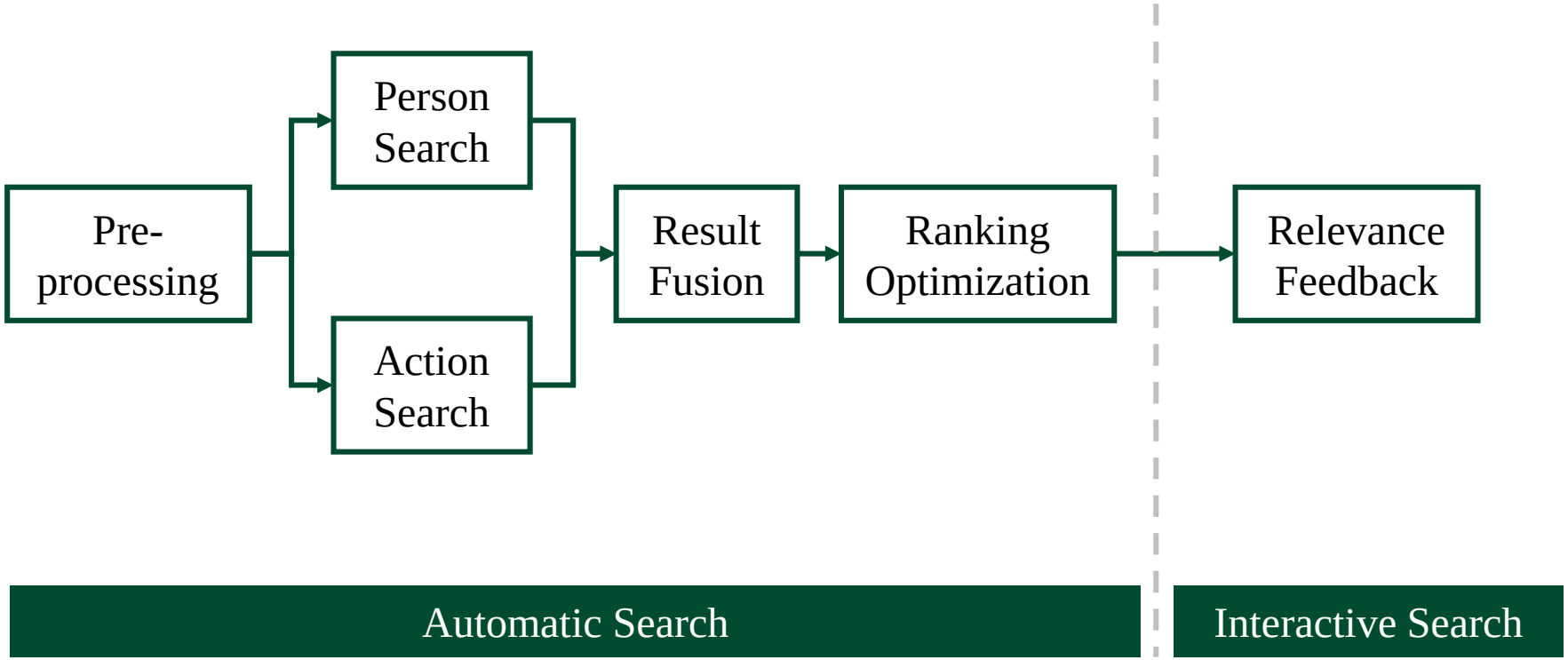
20 topics in TRECVID 2021 INS task

Outline

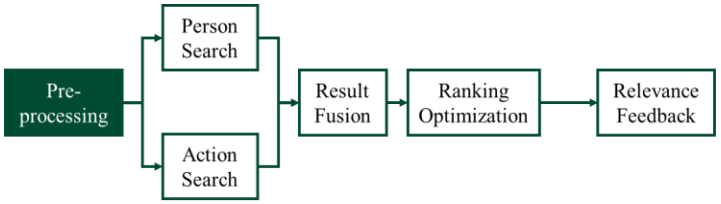
- Introduction
- Approach
- Results
- Conclusions

Approach

- Framework



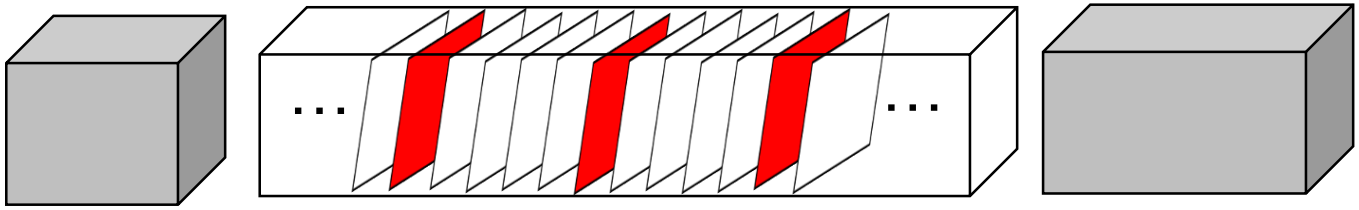
Approach



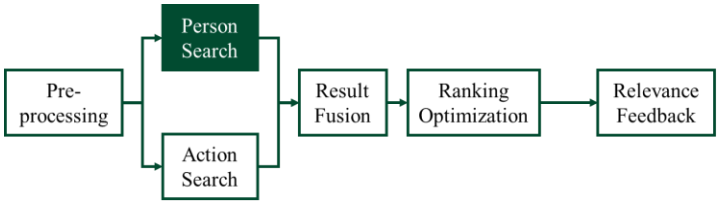
- Step 1: Pre-processing
  - Shot boundary detection (official results)

0	shot0_1	T00:00:00:00F25	T00:00:00:02F25
0	shot0_2	T00:00:00:03F25	T00:00:00:10F25
0	shot0_3	T00:00:00:11F25	T00:00:24:14F25
0	shot0_4	T00:00:24:15F25	T00:00:28:19F25
0	shot0_5	T00:00:28:20F25	T00:00:33:13F25
0	shot0_6	T00:00:33:14F25	T00:00:35:09F25
0	shot0_7	T00:00:35:10F25	T00:00:40:16F25
0	shot0_8	T00:00:40:17F25	T00:00:43:04F25
0	shot0_9	T00:00:43:05F25	T00:00:50:14F25
0	shot0_10	T00:00:50:15F25	T00:00:59:00F25

  - Keyframe extraction
    - Take 1 keyframe from every 5 frames within a shot



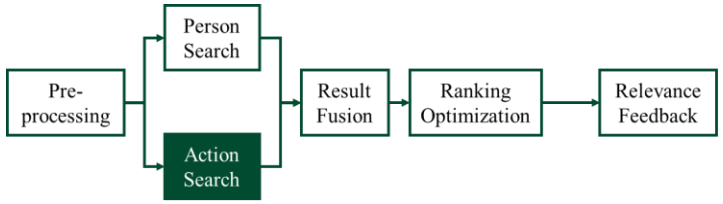
Approach



- Step 2: Person Search
  - Face datasets
    - 2,600+ images corresponding to 191 characters (from bing.com)

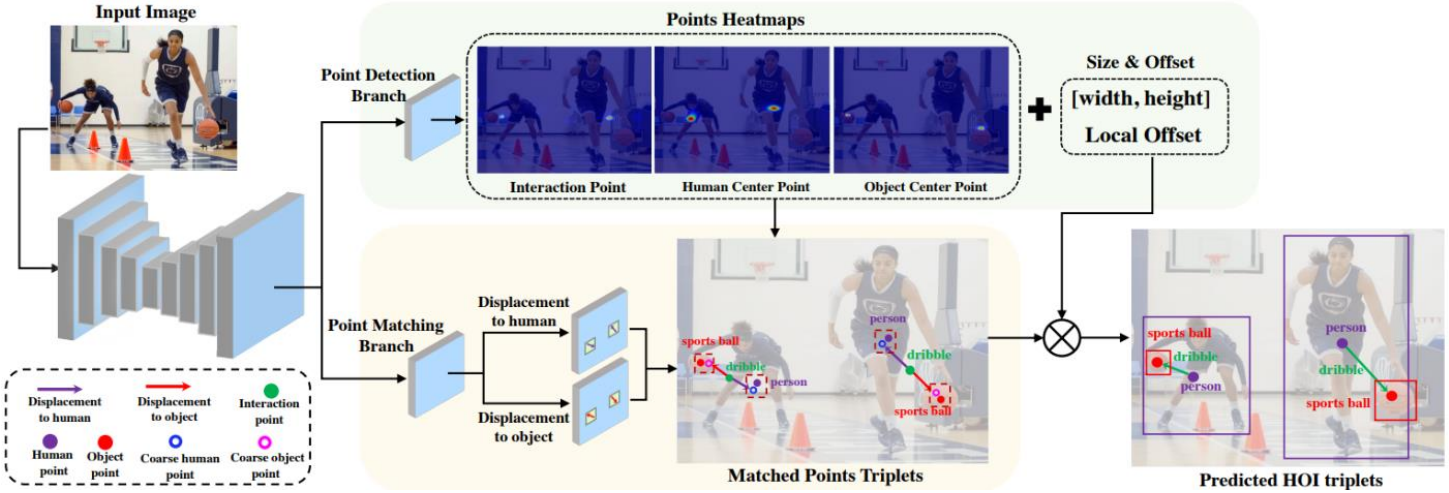
  - Face detection & recognition
    - RetinaFace (CVPR’20), ArcFace (CVPR’19)

Approach



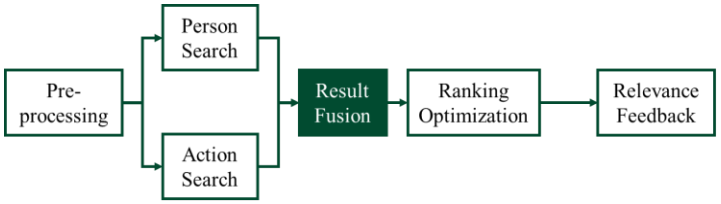
Step 3: Action Search

- Frame level action recognition
  - PPDM (CVPR'20), QPIC (CVPR'21), ASNet (CVPR'21)



- Shot level action recognition
  - TSM (ICCV'19), ACAM (WACV'20)

Approach



Step 4: Result Fusion

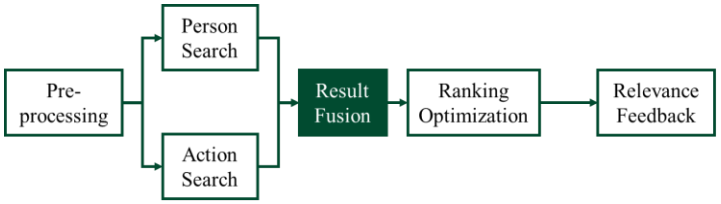
- Identity Inconsistency Problem (IIP)

(a) "Bradley is carrying bag" (b) "Pat is sitting on couch"

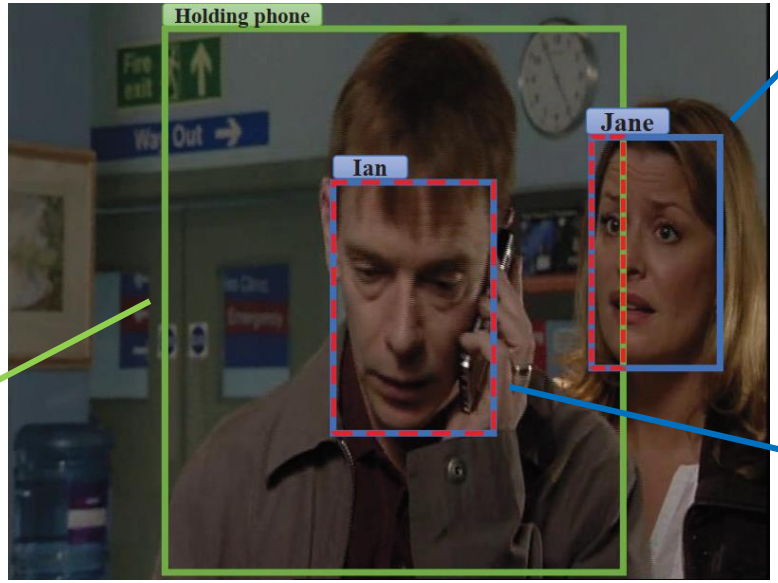
- Statistics
  - IIP causes 23.44% and 22.35% errors in 2019-2020 INS

Jingyao Yang, Chao Liang, Yanrui Niu, Baojin Huang, and Zhongyuan Wang. A spatio-temporal identity verification method for person-action instance search in movies. *arXiv preprint arXiv:2111.00228*, 2021.

Approach



- Step 4: Result Fusion
  - Identity Consistency Verification (ICV)



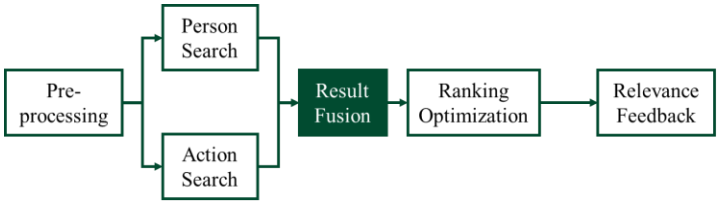
i=1: Jane's face

j=1: Holding phone

i=2: Ian's face

$$c_{i,j} = \frac{\textbf{Intersection}\left(Box_i^{face},\ Box_j^{action}\right)}{\textbf{Area}\left(Box_i^{face}\right)}$$

Approach



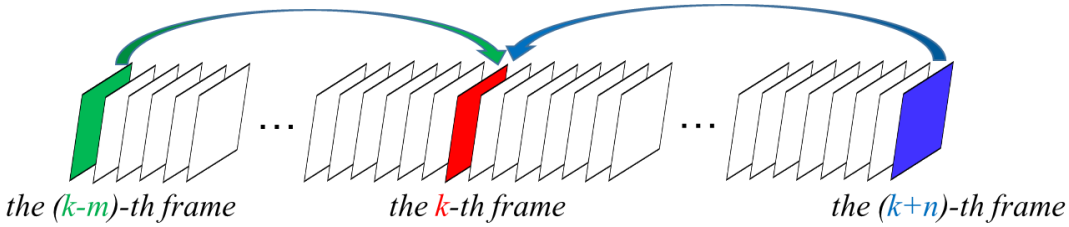
■ Step 4: Result Fusion

- Inter-frame Detection Extension (IDE)

(a) Non-frontal faces (top row) and occluded faces (bottom row)

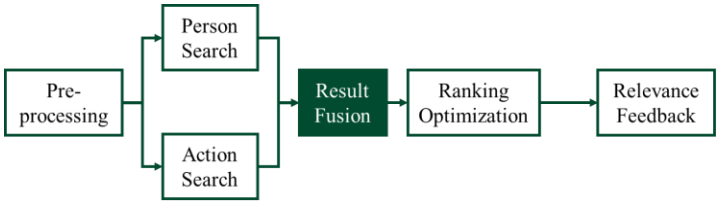
(b) Non-frontal actions (top row) and occluded actions (bottom row)

Linear interpolation:



$$Conf_i^k = \frac{n}{m+n} \times Conf_i^{k-m} + \frac{m}{m+n} \times Conf_i^{k+n}$$

Approach



Step 4: Result Fusion

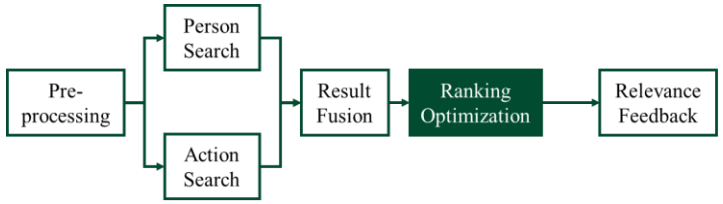
- Fusion strategy: filter-and-rank
  - Step 4.1: use face scores to filter out invalid video shots

$$s_{i,j} = \mathbf{F}_\delta(Conf_i^{face}), \quad \text{where} \quad \mathbf{F}_\delta(x) = \begin{cases} 1, & x \geqslant \delta \\ 0, & x < \delta \end{cases}$$

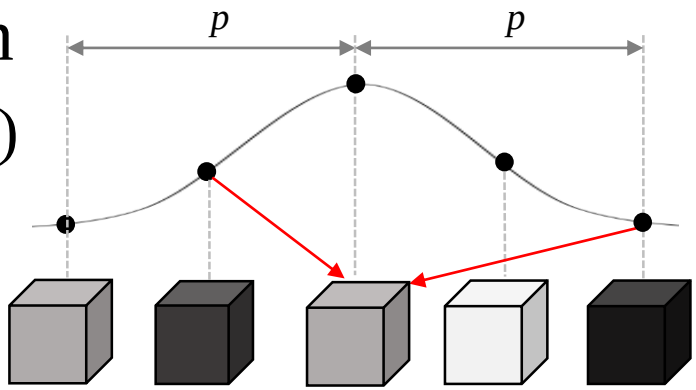
  - Step 4.2: using action scores to rank valid video shots

$$s'_{i,j} = \underbrace{c_{i,j}}_{\text{consistency score}} \times [\underbrace{s_{i,j}}_{\text{person score}} \times \underbrace{Conf_j^{action}}_{\text{action score}}]$$

Approach



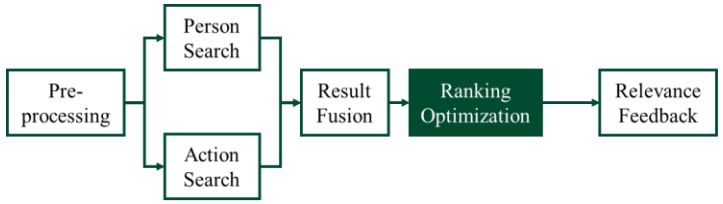
- Step 5: Ranking Optimization
  - Score temporal extension (STE)
    - Video-level score extension



$$s_{i,j}^{k} \leftarrow s_{i,j}^{k} + \theta \sum_{-p<m<p} e^{-\frac{m^2}{\sigma}} \times \max\left(s_{i,j}^{(k+m)} - s_{i,j}^{k},\ 0\right)$$

- Unidirectional extension: from high score shots to low score shots
- Exponential decay: the more distant, the less effective
- Range bounded: between the (k-p)-th and the (k+p)-th frames

Approach



- Step 5: Ranking Optimization
  - Ranking Aggregation (RA)： ER (Omega’20)

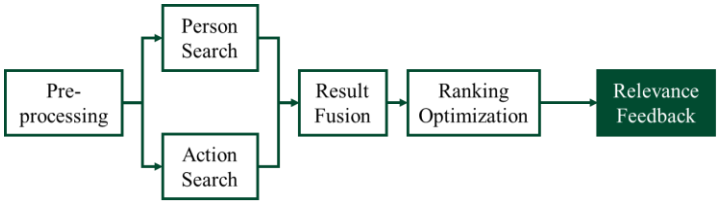
$$\min_{R^*} \frac{1}{2}\sum_{m=1}^{M} g\left(\| R^m - R^* \|_2\right)$$

$$\Leftrightarrow \min_{R^*,\alpha} J(R^*, \alpha) = \sum_{m=1}^{M} \boxed{\alpha_m} \| \boxed{R^m} - \boxed{R^*} \|_2^2 + \psi(\alpha_m)$$

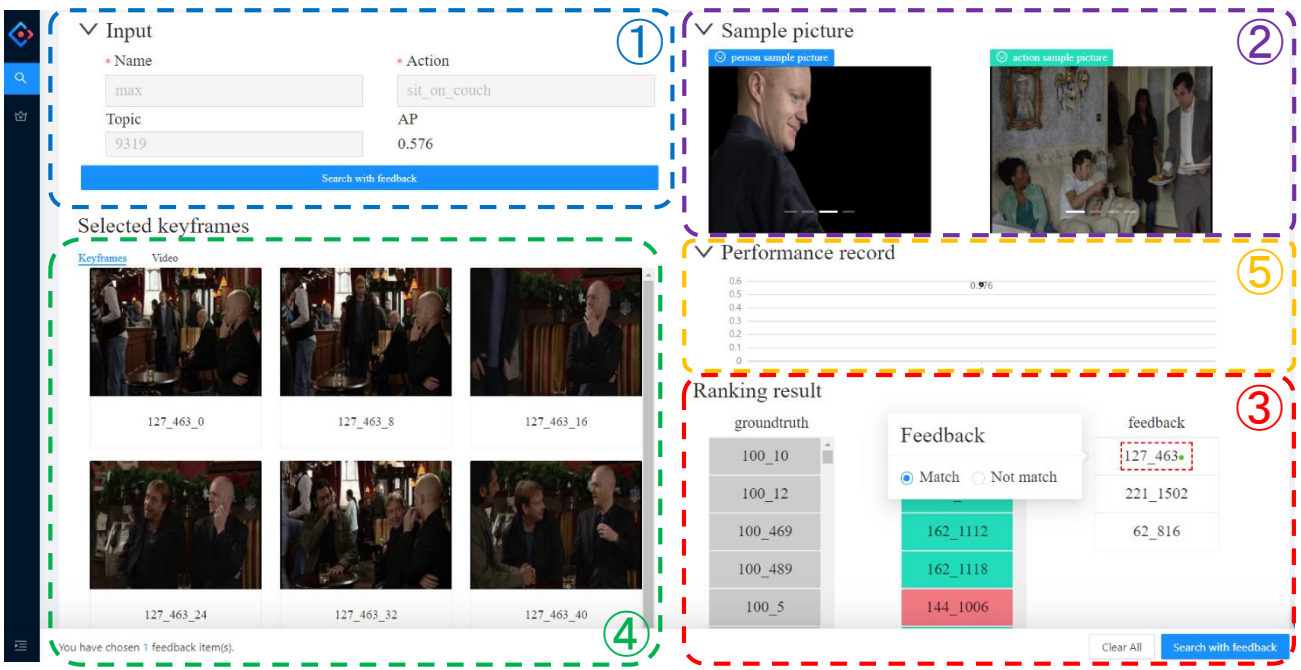
fusion weights　basic rankings　aggregated ranking

- Basic rankings: PPDM, QPIC, ASNet
- Supporting theory: half quadratic programming

Approach



- Step 6: Relevance Feedback
  - GUI for interactive search



① Search topic area

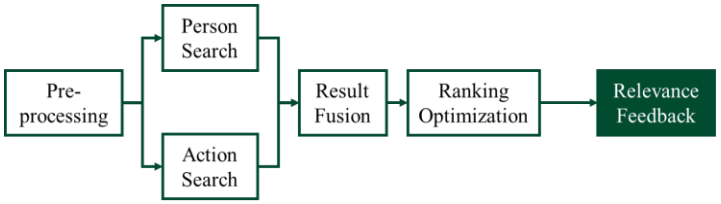
② Query illustration area

③ Ranking & feedback

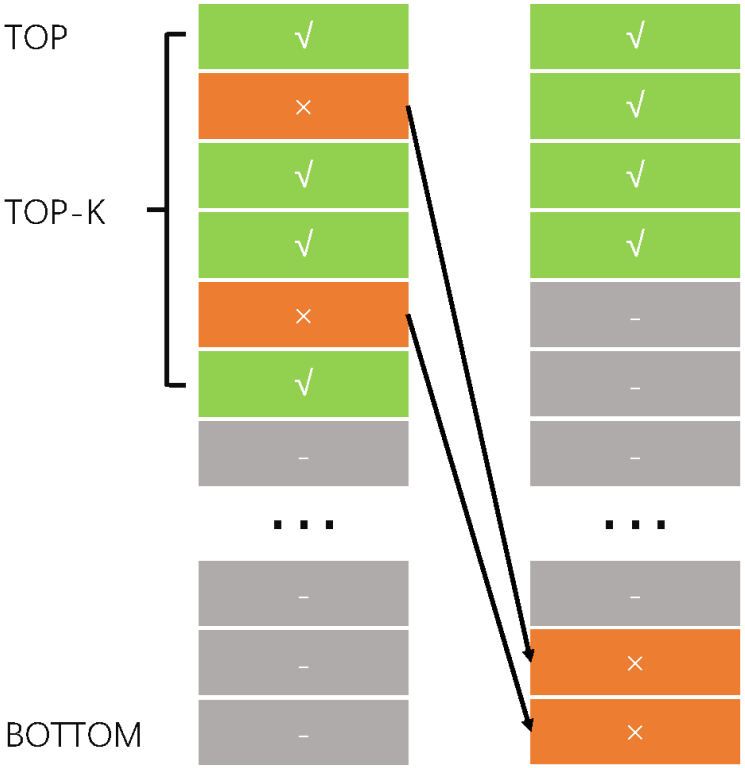
④ Keyframe illustration

⑤ Performance record

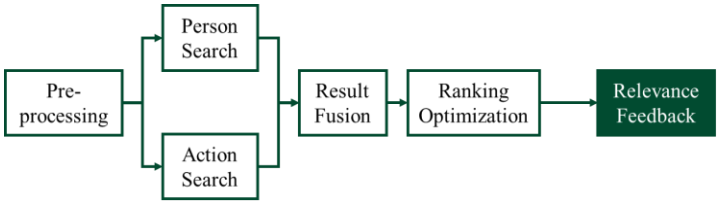
Approach



- Step 6: Relevance Feedback
  - Top-K feedback
    - Browse from top to bottom
    - Feedback on unlabeled items
  - Two acceleration schemes
    - Only positive feedback
    - Only negative feedback

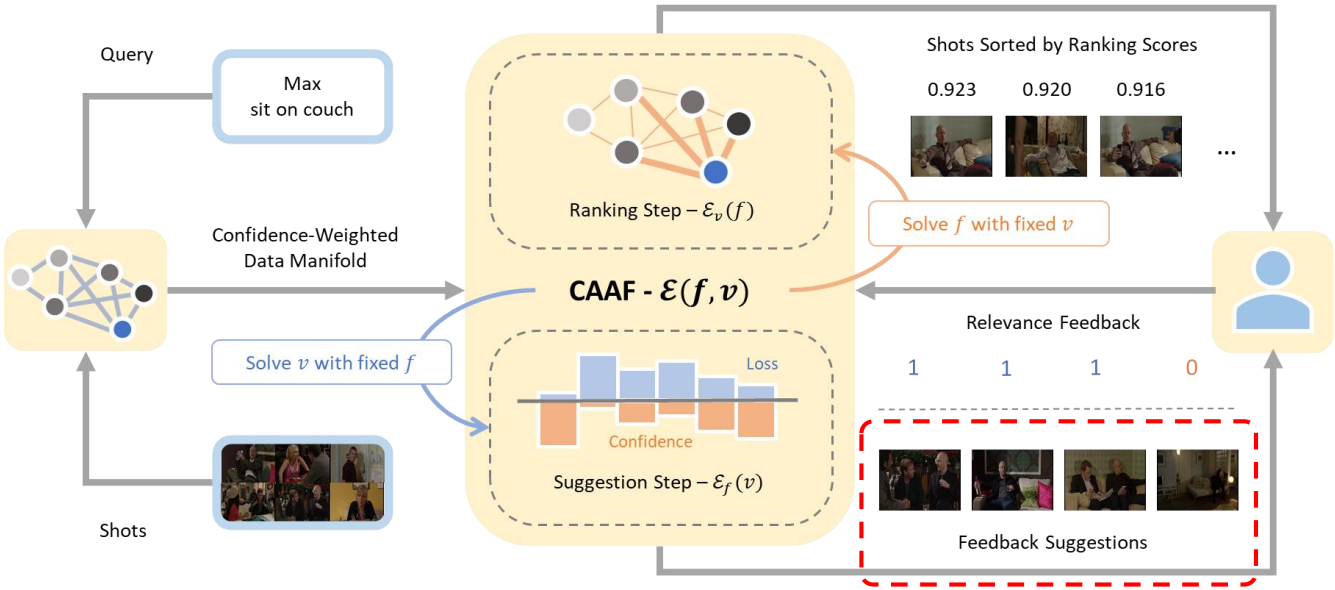


Approach



■ Step 6: Relevance Feedback

- Confidence-aware active feedback (CAAF)



$$\min_{\boldsymbol{f},\,\boldsymbol{v}} \ \mathcal{E}(\boldsymbol{f},\,\boldsymbol{v}) = \mathcal{L}(\boldsymbol{f},\,\boldsymbol{v}) + \mathcal{R}(\boldsymbol{v})$$

Yue Zhang, Chao Liang, and Longxiang Jiang. Confidence-aware active feedback for efficient instance search. arXiv preprint arXiv:2110.12255, 2021

Outline

- Introduction
- Approach
- Results
- Conclusion

Results

- Strategy for Automatic Search
  - 14 topics retrieved by frame level action recognition

<table>
<tr><th colspan="3">RUN ID</th><th>F_2</th><th>F_4</th><th>F_6</th><th>F_8</th></tr>
<tr><th>Topic ID</th><th>Person</th><th>Action</th><th colspan="4">Action recognition and optimization strategy</th></tr>
<tr><td>9319</td><td>Max</td><td rowspan="3">sit_on_couch</td><td rowspan="12">ER[PPDM(IDE+ICV+STE), QPIC(ICV+STE), ASNet(ICV+STE)]</td><td rowspan="6">QPIC
(ICV+STE)</td><td rowspan="6">QPIC
(ICV+STE)</td><td rowspan="6">QPIC
(ICV)</td></tr>
<tr><td>9320</td><td>Stacey</td></tr>
<tr><td>9321</td><td>Peggy</td></tr>
<tr><td>9322</td><td>Stacey</td><td rowspan="3">holding_glass</td></tr>
<tr><td>9323</td><td>Bradley</td></tr>
<tr><td>9324</td><td>Shirley</td></tr>
<tr><td>9325</td><td>Bradley</td><td rowspan="3">holding_phone</td><td rowspan="6">PPDM
（IDE+ICV+STE）</td><td rowspan="6">PPDM
（IDE+ICV+STE）</td><td rowspan="6">PPDM
（IDE+ICV）</td></tr>
<tr><td>9326</td><td>Shirley</td></tr>
<tr><td>9327</td><td>Peggy</td></tr>
<tr><td>9328</td><td>Max</td><td rowspan="2">carrying_bag</td></tr>
<tr><td>9329</td><td>Peggy</td></tr>
<tr><td>9330</td><td>Pat</td><td rowspan="3">holding_paper</td></tr>
<tr><td>9331</td><td>Shirley</td></tr>
<tr><td>9332</td><td>Peggy</td></tr>
</table>

Results

Strategy for Automatic Search

- 6 topics retrieved by shot level action recognition

RUN ID			F_2	F_4	F_6	F_8
Topic ID	**Person**	**Action**	**Action recognition and optimization strategy**			
9333	Max	kissing	TSM	TSM	TSM (STE)	TSM
9334	Stacey					
9335	Bradley	open_door_enter	ACAM (IDE+ICV)	ACAM (IDE+ICV)	ACAM (IDE+ICV)	ACAM (IDE+ICV)
9336	Pat					
9337	Max	holding_cloth	TSM	TSM	TSM	TSM
9338	Stacey					

Strategy for Interactive Search

I_1	I_3	I_5	I_7
F_2 + TopK	F_4 + CAAF	F_4 + TopK	F_8 + CAAF

Results

Overall Results

Type	Run ID	Relation	mAP	Strategy
Automatic	F_2	--	**0.435**	A + F + S + R
	F_6	--	0.418	A + F + S
	F_4	--	0.418	A + F + S (w/o STE on kissing)
	F_8	--	0.395	A + F
Interactive	I_1	F_2 + Top-K	**0.465**	A + F + S + R + $I_{Top\text{-}K}$
	I_5	F_4 + Top-K	0.460	A + F + S + $I_{Top\text{-}K}$
	I_3	F_4 + CAAF	0.459	A + F + S + I_{CAAF}
	I_7	F_8 + CAAF	0.443	A + F + I_{CAAF}

Abbreviation	Description	Abbreviation	Description
A	Action recognition	R	Ranking aggregation
F	Face recognition	$I_{Top\text{-}K}$	Top-K feedback
S	Score temporal extension	I_{CAAF}	CAAF feedback

Results

- Result Fusion
  - ICV & IDE

Topic ID			9319-9332			9335-9336
Original	**IDE**	**ICV**	**PPDM**	**QPIC**	**ASNet**	**ACAM**
√			39.63	36.75	39.78	6.56
√	√		39.15	36.04	38.98	6.59
√		√	**43.36**	**41.94**	**43.36**	8.13
√	√	√	43.21	41.74	42.83	**8.24**

  - Discussion
    - ICV is always effective, while IDE is not
    - IDE is slightly effective when original performance is bad

Results

- Ranking Optimization
  - STE

Topic ID		9319-9332		
Basic*	**STE**	**PPDM**	**QPIC**	**ASNet**
√		43.21	41.94	43.36
√	√	**47.51**	**43.89**	**44.83**

  - Discussion
    - STE is effective
    - Improvement varies among different ranking methods

Remark: In F_2 setting, basic ranking of PPDM is obtained by Original+IDE+ICV, basic ranking of QPIC and ASNet is obtained by Original+ICV (without IDE).

Results

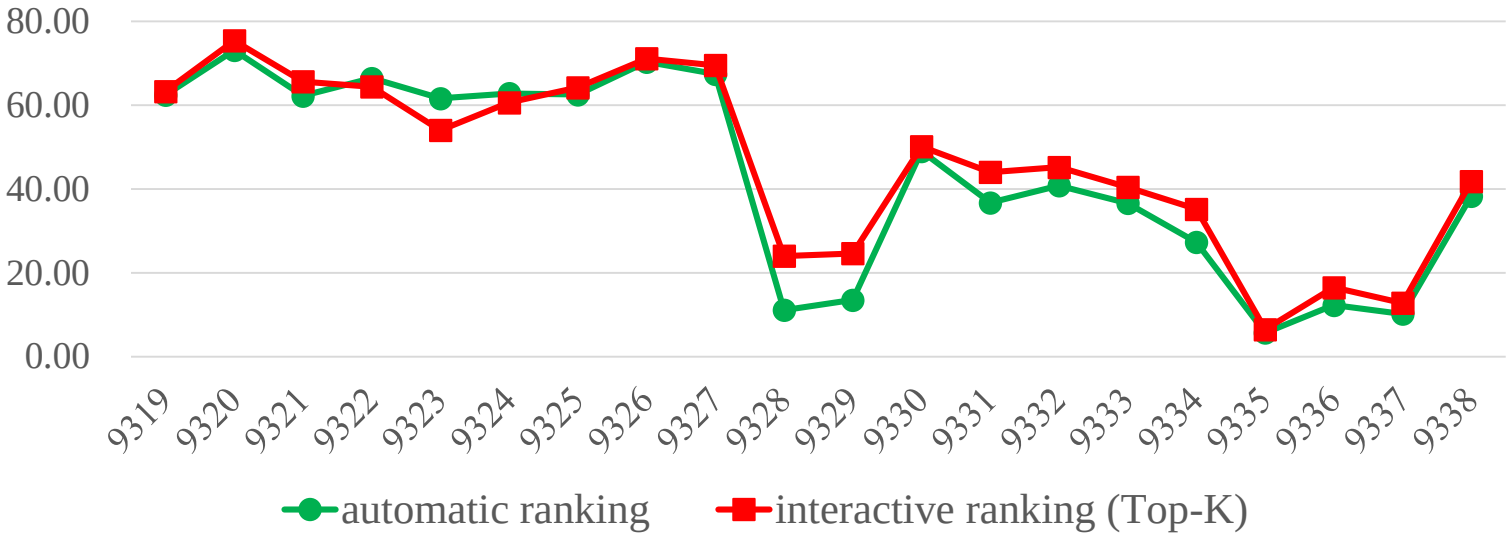
- Ranking Optimization
  - ER

9319-9332			
PPDM	**QPIC**	**ASNet**	**ER [PPDM+QPIC+ASNet]**
47.51	43.89	44.83	**52.84**

  - Discussion
    - ER significantly improves ranking performance of basic rankers

Results

- Relevance Feedback
  - Top-K (I_1)



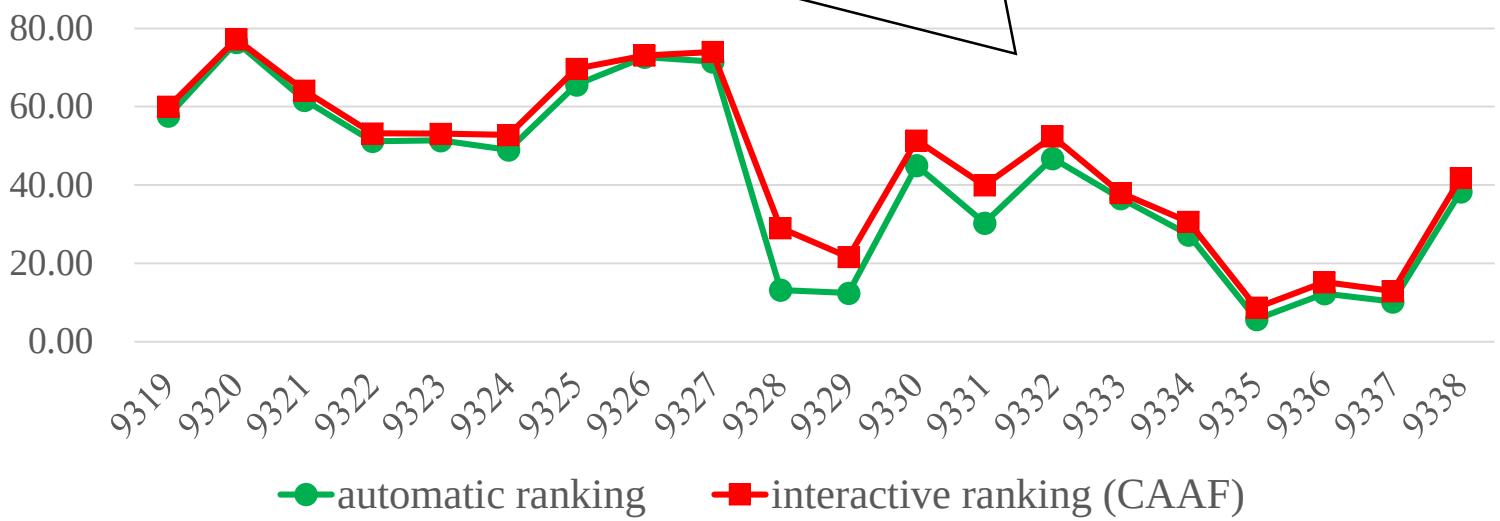
  - Discussion
    - Top-K scheme is effective
    - User may make mistakes in the relevance feedback, e.g., 9323

Results

- Relevance Feedback
  - CAAF (I_3)

Top-K: mAP(interactive-automatic) = 3.60

CAAF: mAP(interactive-automatic) = 4.45



  - Discussion
    - CAAF is effective
    - CAAF's improvement is larger than that of Top-K

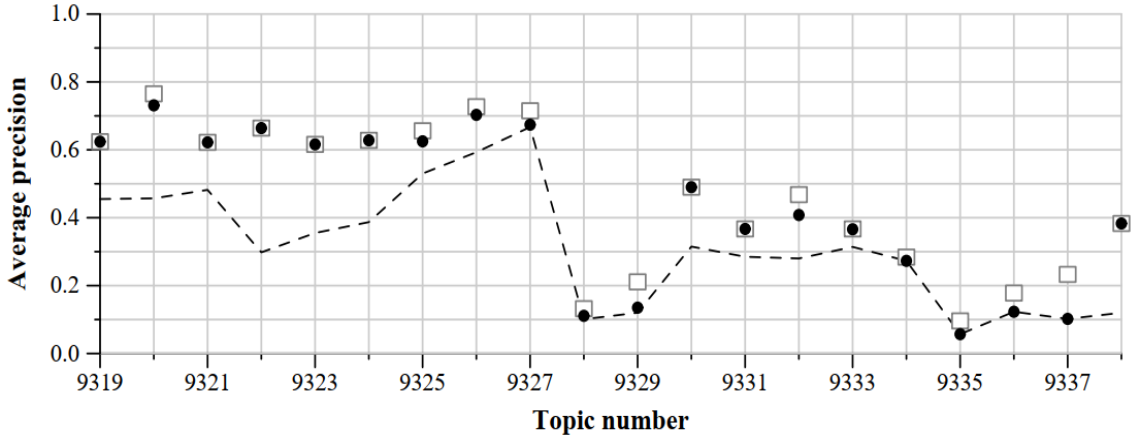
Results

- Relevance Feedback
  - Top-K (I_5) vs. CAAF (I_3)
    - Top-K is slightly better than CAAF, i.e., 46.0 vs. 45.9, within 5-min feedback
    - CAAF achieves nearly the same performance as Top-K with much less feedbacks, i.e., 78.4 vs. 105.4
    - CAAF remarkably surpasses Top-K given the same quantity of feedback items, i.e., 51.9 vs. 46.0

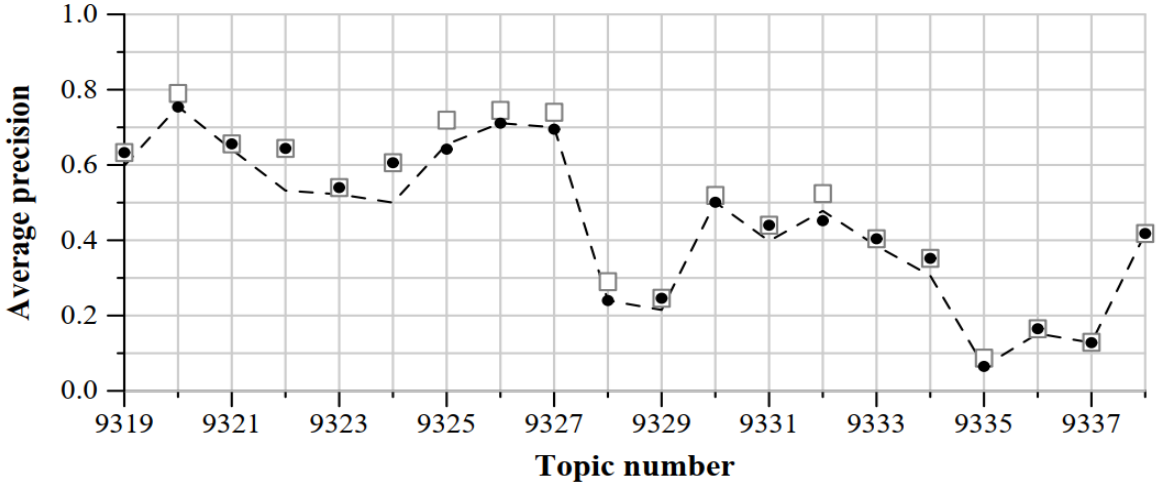
Topic	Official				Ours	
	Auto2+TopK		Auto2+CAAF		Auto2+CAAF*	
	#FB	AP	#FB	AP	#FB	AP
9319	176	59.8	75	60.0	176	70.7
9320	169	79.0	67	77.2	169	80.2
9321	182	65.3	80	64.1	182	72.1
9322	97	56.4	60	53.2	97	59.1
9323	90	51.5	66	53.1	90	57.8
9324	91	49.4	61	52.8	91	56.0
9325	120	71.9	77	69.7	120	75.3
9326	138	74.5	99	73.1	138	78.4
9327	130	74.0	90	74.0	130	75.7
9328	98	20.9	86	29.0	98	34.3
9329	114	21.5	90	21.6	114	24.5
9330	97	51.9	73	51.4	97	55.2
9331	104	39.8	80	39.9	104	45.9
9332	116	51.6	80	52.4	116	60.2
9333	77	40.4	82	38.0	77	45.9
9334	79	35.2	83	30.6	79	45.1
9335	60	6.5	80	8.7	60	14.0
9336	67	16.5	96	15.2	67	24.0
9337	50	12.8	72	12.9	50	15.9
9338	53	41.8	70	41.8	53	46.7
Mean	105.4	46.0	78.4	45.9	105.4	51.9

Results

- Best Results
  - Automatic
    - F_2



  - Interactive
    - I_1



Outline

- Introduction
- Approach
- Results
- Conclusions

Conclusion

- P-A INS
  - A Divide-and-Fuse-and-Optimization framework
- IIP
  - Check identity consistency is important for P-A INS
  - ICV is robustly effective, while IDE is not
- Ranking Optimization is important
  - STE and ER are both proved effective
- Interaction methods are helpful
  - Top-K and CAAF are both provided effective

Thanks for your time!

A team's work presented by

Yanrui Niu†, Jingyao Yang†, Ankang Lu†, Baojin Huang†,
Yue Zhang, Ji Huang, Shishi Wen, Dongshu Xu, Chao Liang*, Zhongyuan Wang*, Jun Chen*
(† indicates equal contribution, * indicates corresponding authors)

Hubei Key Laboratory of Multimedia and Network Communication Engineering
National Engineering Center for Multimedia Software
School of Computer Science, Wuhan University

(reported on December 10, 2021, revised on December 25, 2021)