Activity and Relationship Modeling Driven Weakly Supervised Object Detection

Yinlin Li[1], Yang Qian[1], Xu Yang[1], Zhang Yuren[2]

1. State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences
2. ByteDance

Motivation

Configuration of human and object are similar in same activity, and joint modeling of human, active object and activity could leverage the recognition of them.

Method

Step 1. Proposal Selection Based on Class Activation Map

- **The most salient active human box:**

$$b_{h*} = \underset{b_h}{\mathrm{argmax}}(\alpha_1 S_h + \alpha_2 \overline{\textstyle\sum_{c_a}\sum_{p_x \in b_h} M_{c_a}(p_x)} + \alpha_3 \overline{\textstyle\sum_{c_a}\sum_{p_x \in b_{h^*}} M_{c_0}(p_x)})$$

- **The weighted score for each candidated object box:**

$$s_{b_0} = \beta_1 s_0 + \beta_2 \overline{\textstyle\sum_{p_x \in b_0} M_{c_0}(p_x)} + \beta_3 \overline{\textstyle\sum_{c_a}\sum_{p_x \in b_{0^*}} M_{c_a}(p_x)}$$

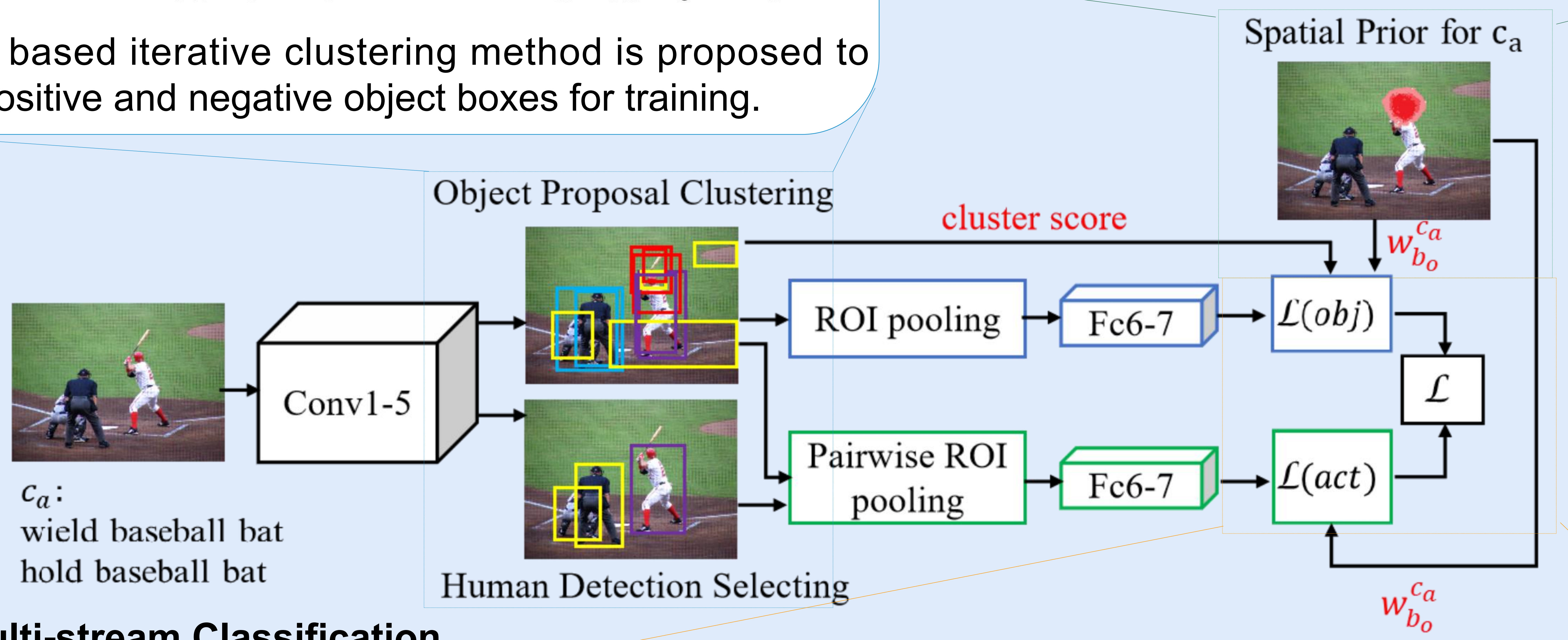
A score based iterative clustering method is proposed to select positive and negative object boxes for training.

Step 2. Spatial Relationship Modeling

- **Multiple geometric relations:**

$$r_{(b_o,b'_h)} = [\frac{x_{b_o} - x_{b'_h}}{\sqrt{W_{b'_h} H_{b'_h}}}, \frac{y_{b_o} - y_{b'_h}}{\sqrt{W_{b'_h} H_{b'_h}}}, \sqrt{\frac{W_{b_o} H_{b_o}}{W_{b'_h} H_{b'_h}}}, IoU(b_o, b'_h), \frac{W_{b_o}}{H_{b_o}}, \frac{W_{b'_h}}{H_{b'_h}}],$$

- **Spatial gaussian prior of object and action:** $w_{b_o}^{c_a} = \mathcal{N}_{(\mu_{c_a}, \sigma_{c_a})}(r_{(b_o,b'_h)})$



Step 3. Multi-stream Classification

- **Spatial prior weighted object classification loss:**

$$\mathcal{L}(b_o) = -\sum_{c_o} \hat{s}_{b_o} log(g(c_o|b_o)) + (1 - \hat{s}_{b_o})(1 - log(g(c_o|b_o)))$$

$$\mathcal{L}(obj) = \frac{1}{n_{b_o}} \frac{1}{n_{c_a}} \sum_{b_o} \sum_{c_a} w_{b_o}^{c_a} \mathcal{L}(b_o)$$

- **Spatial prior weighted activity classification loss:**

$$\mathcal{L}(act) = -\frac{1}{n_{b_o}} \frac{1}{n_{c_a}} \sum_{b_o} \sum_{c_a} w_{b_o}^{c_a} [y_{c_a} log(h(c_a|b_o, b'_h)) + (1 - y_{c_a})(1 - log(h(c_a|b_o, b'_h)))]$$

- **Integrated loss function:** $\mathcal{L} = \lambda_o \mathcal{L}(obj) + \lambda_a \mathcal{L}(act)$

Results and Conclusion

Comparison of active human selection methods

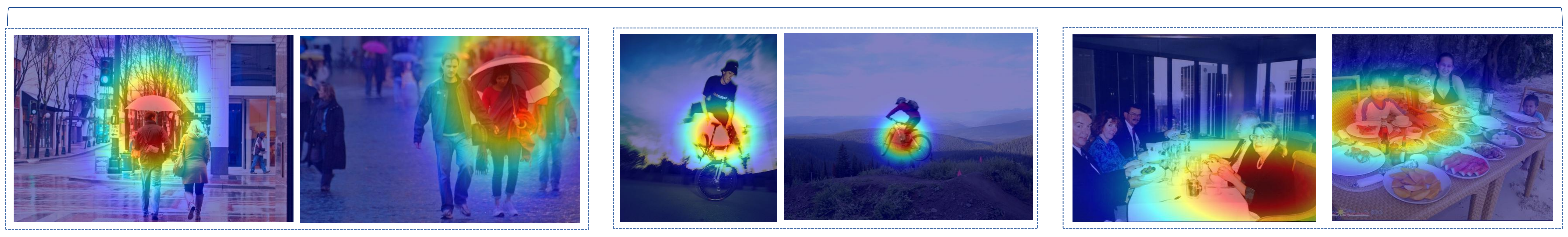
Methods	all data	sampled multi-person data
Method in [18]	73.28	54.16
Ours	81.24	77.08

Comparison of object proposal selection methods

Methods	recall(%)	precision(%)
700 proposals	73.82	0.72
1200 proposals	81.68	0.72
our method	54.21	24.12

Active human selection examples (red: our method)

Spatial prior learning results of different activities



Object detection AP results on HICO-DET

Methods	mAP(%)
R*CNN [32]	2.15
WSDDN [4]	3.27
PCL [11]	3.62
PCL + prior	4.19
ASDNN	5.39
Ours without gaussian prior	7.82
Ours	8.11

A weakly supervised object detection method based on activity class level supervision is proposed, which has three highlights:
1) Active human and candidate object proposals are learned, filtered and clustered with higher accuracy/precision;
b) Spatial Gaussian prior is modeling based on multiple geometric relations to improve the localization precision of object;
c) object and activity classifications are integrated together, and the final result outperforms the SoTA methods.