JRDB-Act: A Large-scale Dataset for Spatio-temporal Action, Social Group and Activity Detection

Supplementary Material

Mahsa Ehsanpour[1], Fatemeh Saleh[2*], Silvio Savarese[3], Ian Reid[1], Hamid Rezatofighi[4]
[1]The University of Adelaide, [2]Samsung AI Center, [3]Stanford University, [4]Monash University
https://jrdb.erc.monash.edu/

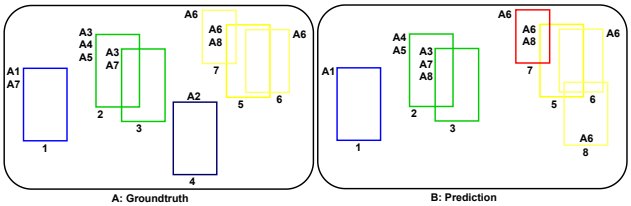

Figure 1. An example of the groundtruth and prediction scenarios for the three sub-tasks of individual action, social group and social activity detection. Matched groundtruth and detected bounding boxes are shown with similar number. The set of actions for each box is indicated by Ai next to it.

In this supplementary material, we explain details of the sections that refer readers to the supp.material in the original paper.

1. Metric and Evaluation

We evaluate three sub-tasks on JRDB-Act namely, individual action, social group and social activity detection. We utilize the widely used Mean Average Precision (mAP) on the key-frames of test-set. At inference, for each detected bounding box in the key-frame, our model predicts a set of individual action labels, a social group ID and we infer a set of social activity labels for that box by utilizing its social group ID and the predicted individual action labels of that group's members. The inferred social activity labels for all the bounding boxes of a social group, would be the common individual action labels of the members of that group, *i.e*. the actions performed by two or more people in the same group. Note that for singleton groups (groups with one member), the social activity labels is identical to the person's individual actions. To explain the evaluation strategy of the three tasks, we show an example in Fig. 1. We show bounding boxes by rectangles. Each bounding box has at least one individual action indicated by Ai and the color of bounding boxes indicate their social group ID. Consider A and B in Fig. 1 as the groundtruth and the prediction scenarios respectively.

A. Individual Action Evaluation. Individuals' action detection evaluation is similar to the standard practice in [2]. The true positive cases are (box1, A1), (box2, A4), (box2, A5), (box3, A3), (box3, A7), (box5, A6), (box5, A8), (box6, A6), (box7, A6) and the false negative cases are (box1, A7), (box2, A3), (box4, A2) and the false positive cases are (box3, A8), (box8, A6).

B. Social Grouping Evaluation. For the social grouping evaluation, true positive cases are boxes 1,2,3,5,6, false negative case is box 4 and false positive cases are boxes 7,8. Social grouping performance is reported for social groups with 1, 2, 3, 4 and 5 or more members indicated by G1, G2, G3, G4 and G5$^+$ AP in Tab. [2] and Tab. [3] of the paper. Overall AP in these tables is the average of G1-G5$^+$.

C. Social Activity Evaluation. For the social activity evaluation, the inferred groundtruth social activity labels for the blue group is A1, A7, for the green group is A3, for the navy group is A2 and for the yellow group is A6. Similarly, the predicted social activity labels for the blue group is A1, for the yellow group is A6 and for the red group is A6. For the green group we consider no predicted social activity label since none of the individual actions happening in that group is done by at least two people in the group. We assign the groundtruth and predicted per-group social activity labels to the members of that group. We evaluate social activity detection task in two ways. G-Act mAP1 evaluate the task by not considering the predicted social groups similar to the the individual action detection evaluation. In this scenario, true positive cases are (box1, A1), (box5, A6), (box6, A6) and **(box7, A6)** although the social group of box 7 is wrongly predicted. False negative cases are (box1, A7), (box2, A3), (box3, A3), (box4, A2) and the false positive cases are (box8, A6). On the other hand, G-Act mAP2 evaluate the task by considering the predicted social groups. In this scenario, **(box7, A6)** is a false positive since its predicted group membership is not correct. Obviously, G-Act mAP2 is a stricter metric than G-Act mAP1 and clarifies the reason of lower performance in G-Act mAP2 compared to

*Work done while at the Australian National University (ANU).

G-Act mAP1 in Tab. [3] and Tab. [4] of the paper.

2. Eigen-value based loss Proof.

As stated in **Learning Social Group Formation** of Sec. [4] in the paper, the number of connected components (social groups) in the groundtruth matrix A is equal to the number of zero eigenvalues of its laplacian matrix L. Thus, we want the laplacian matrix of A_θ denoted by L_θ to have the same number of zero eigenvalues as in L. If e_θ is an eigenvector of $\mathrm{L}_\theta^T\mathrm{L}_\theta$ (in order to ensure that the matrix is symmetric) with the eigenvalue λ, it satisfies $\mathrm{L}_\theta^T\mathrm{L}_\theta e_\theta = \lambda e_\theta$. Since $e_\theta^T e_\theta = 1$ (eigenvectors have unit-norm), multiplying both sides of the equation by e_θ^T yields $e_\theta^T\mathrm{L}_\theta^T\mathrm{L}_\theta e = \lambda$ and we want to consider eigenvalues of zero ($\lambda = 0$). Given the groundtruth eigenvector e corresponding to the zero eigenvalue, we define the loss as

$$L_{eig}(\theta) = e^T\mathrm{L}_\theta^T\mathrm{L}_\theta e; \quad e^T\mathrm{L}_\theta^T\mathrm{L}_\theta e \geq 0 \tag{1}$$

To avoid the trivial solution $\mathrm{L}_\theta = 0$, a second term is added to maximize the projection of data along the directions orthogonal to e.

$$L_{eig}(\theta) = e^T\mathrm{L}_\theta^T\mathrm{L}_\theta e - \alpha tr(\bar{\mathrm{L}}_\theta^T\bar{\mathrm{L}}_\theta) \tag{2}$$

and finally for numerical stability the second term is bounded in the range $[0, 1]$ as

$$L_{eig}(\theta) = e^T\mathrm{L}_\theta^T\mathrm{L}_\theta e + \alpha \exp(-\beta tr(\bar{\mathrm{L}}_\theta^T\bar{\mathrm{L}}_\theta)) \tag{3}$$

This fully differentiable, eigendecomposition-free loss allows us to avoid performing eigendecomposition which suffers from the numerical instabilities of analytical differentiation.

3. JRDB-Act Action partitions.

As stated in **Learning Actions** of Sec. [4] in the paper, to improve the performance of individuals action detection at the presence of highly unbalanced action label distribution, we propsoe to utilise partitioning and balancing action loss functions based on the occurring frequency of action classes in the dataset. We utilise 3 cross entropy losses for 3 pose-based partitions: **[walking, standing, sitting], [cycling, going upstairs, bending], [going downstairs, skating, scootering, running]** and one binary cross entropy loss to learn whether there exists any interaction-based action and 3 more binary cross entropy losses for interaction-based partitions: **[holding sth, listening to someone, talking to someone], [looking at robot, looking into sth, looking at sth, typing, interaction with door, eating sth], [talking on the phone, reading, pointing at sth, pushing, greeting gestures]**. Action labels are divided into disjoint partitions such that the occurring frequency of the most frequent action class is no more than 10 compared to the least frequent class in that partition.

4. Implementation Details.

Our model's backbone $f_\theta(x)$ is obtained from [1] and here we elaborate its imp. detail. We uses an I3D feature extractor which is initialized with Kintetics-400 [3] pre-trained model. We utilize ROI-Align with crop size of 5×5 on extracted feature-map from I3D. We perform self-attention on each individuals' feature map with query, key and value being different linear projections of individuals' feature map with output sizes being $1/8, 1/8, 1$ of the input size. Individuals' feature maps obtained from the self attention module are fed into a single-layer, multi-head graph attention module with 8 heads with input, hidden and output dimention size of 1024 and droupout probability of 0.5 and $\alpha = 0.2$ [4]. The rest of the imp. detail is included in the paper.

References

[1] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Saleh, Javen Shi, Ian Reid, and Hamid Rezatofighi. Joint learning of social groups, individuals action and sub-group activities in videos. In *ECCV*, 2020. 2

[2] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, pages 6047–6056, 2018. 1

[3] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[4] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2017. 2