Rebuttal for *Binary Coding for Partial Action Analysis with Limited Observation Ratios*

We are grateful for all the respectful reviewers' suggestions, and greatly encouraged by the positive comments, e.g. the idea is novel and well formulated; significant superiority to state-of-the-arts in solid experiments; the rationale and contributions are clearly written; the first work using hashing for partial action analysis in videos.

1. To Reviewer 1: Thanks a lot for your significant support.

2. To Reviewer 2: We are greatly encouraged by your insightful comments, e.g. novel idea and excellent results. Thanks a lot for all the appreciation.

On the positive results: As stated in L557-564, all partial/full actions adopt 'C3D' to obtain representations. Our PRBC and the alternative ones are all evaluated based on the same 4096-d C3D descriptors. Therefore, the comparison with alternative approaches is fair and the positive results are indeed due to the proposed strategy. The large enhancement owes to the following advantages of PRBC:

1) 'Feature Reconstruction' can transfer crucial information from full to partial actions. Though Eq. 1 seems somewhat restrictive, this lease squares-style function has shown its effectiveness in various tasks. Besides, as shown in Fig. 6 (a) and (b), though the input descriptors are highly discriminative, there is still a large gap between the distributions of partial and full actions. So if we directly apply the model trained on full actions to partial ones, there will be less effective results (i.e. 'CF' rows in Table 1/2). After reconstruction, the results can be improved by a large margin, e.g. 5-10 times better than 'CF', as shown in 'CF+R' rows.

2) 'Binary Coding' can further enhance the discrimination of PRBC. Eq. 2 and 3 ensure more related actions to be projected to more similar codes. This can be seen from the performance gain of PRBC over 'CF+R' since PRBC is essentially equivalent to 'CF+R'+'Binary Coding'.

3) As Fig. 6 (c) and (d) show, more separable embeddings are learned and partial/full actions share similar embeddings after PRBC. This further proves its effectiveness.

On the extension to longer video actions: PRBC is basically formulated regardless of the video length, as Eq. 6 involves no variable w.r.t. video length. Practically, we can extend PRBC to deal with longer videos in these ways:

1) We could utilize an additional step to handle longer videos. Specifically, in the training stage, we employ some action proposal [1] or 'actionness' detection [2] methods to discover the meaningful parts involving actions. Then PRBC can be learned based on the cropped full action videos that include more relevant parts. Accordingly, when a partial video comes, we first evaluate the 'actionness' of this video. If it tends to be a part of an action, we then perform further action analysis. Otherwise, we simply treat it as the non-relevant part without any further analysis.

2) In our experiments, we found that using more temporal segments, i.e. larger $\mathbf{X}$ (L249-250), to train the reconstruction function could result in more accurate results. Hence, in terms of longer videos, we could alternatively select much more segments to learn a more effective reconstruction function (if given enough memory resource).

On the per-class analysis: Thank you and we will add a confusion table for the per-class results in the final version.

3. To Reviewer 3: Thank you for your comments.

On the problem definition: This definition is commonly adopted in previous works [4, 15, 19]. As clearly stated in L78-80 and shown in Fig. 1 (c), a 'partial action' means a shorter **'temporal'** segment of the complete action, i.e. we consider partial actions that miss some temporal information. Hence, a video with partly occluded actions is apparently out of the scope of this paper.

On the arbitrary ORs: Existing methods assume ORs of partial actions are known during testing (L115-118). In our setting, we need not know ORs (L576-580) and use 'arbitrary ORs' to claim this point. You must have been confused of the word 'arbitrary', thus had some misunderstandings. We will instead use 'unknown ORs' for disambiguation in the final version.

On the additional questions: 1) Of course we can. Due to space limit, we theoretically analyze the convergence (L378-383). In practice, PRBC can converge within 3$\sim$5 iterations (L501-502). 2) We have already utilized some partial actions to learn the hashing representation (L647 and L665-667), which exactly corresponds to your suggested setting. We can see this strategy cannot achieve the satisfactory performance. 3) The results of other methods are the best ones based on the optimized parameters (L668-670). 4) We can confirm all methods use the exact same training/testing videos. We strictly follow the standard 3 splits on HMDB51/UCF101 (L730-732), and the standard setting on UT-Interaction (L790). 5) L249-254 visually depicts how to learn the reconstruction function, i.e. we utilize multiple short temporal segments to approximate the corresponding full action video. 6) '-1' and '1' are widely used values for affinity parameters [23, 29, 32]. Theoretically, s could have other values instead of '1.5'. Actually, we tried to use different values from 1 to 3 with step-size 0.5 and found that '1.5' was the best choice.

References

[1] G. Gkioxari and J. Malik. Finding action tubes. In *CVPR*, 2015. 1

[2] L. Wang, Y. Qiao, X. Tang, and L. V. Gool. Actionness estimation using hybrid fully convolutional networks. In *CVPR*, 2016. 1