Motion Adaptive Pose Estimation from Compressed Videos (Supplementary Materials)

Zhipeng Fan[1] Jun Liu[2*] Yao Wang[1]
Tandon School of Engineering, New York University, Brooklyn NY, USA[1]
Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore[2]
zf606@nyu.edu jun_liu@sutd.edu.sg yw523@nyu.edu

1. Additional Visualization

We further provide additional visualization results for the decisions made by the gate and the pose estimation results in Fig. 1 for (a) Penn Action dataset and (b) Sub-JHMDB dataset. In general, for relatively static GOP, e.g. (i), (ii), (iv) and (v) in Fig. 1, the gate is only activated for a fraction of P-frames, which are often the frames with relatively large motion compensated residuals. Furthermore, for motions captured by the non-activated P-frames, even without any input information from the fully decoded frames, our Motion Adaptive Pose Net could still derive accurate pose sequences, as shown in the first half of the GOP (v). The arm naturally folds along with the moving-up of body joints through the pull-up exercise. This indicates that the motion compensated features could serve as a qualitative proxy for the accurate features that are extracted from the actual decoded frames. As shown in this visualization, the *freely-available motion* could be *efficiently* employed to inject effective motion information for *accurate* pose estimation, which coincides with our motivation to employ those efficient motion warped features for fast pose estimation.

While for GOP contains more pose variations like (iii) and (vi), the macroblock based motion vector could often be less accurate, leading to more motion compensated errors. In this scenario, the gate will be activated more often as indicated in GOP (iii) and (vi). In addition to that, notice our Residual Driven Dynamic Gate activates less often for the first half of the GOP, which corresponds to the setup postures in golf/baseball, while it activates more frequently in the second half, corresponding to the ball striking phase. Given that the ball striking phase often contains faster and more violent motions compared to the setup stage, the macroblock based motion field is often less accurate and therefore the Residual Driven Dynamic Gate determines to extract features from the frames more actively. This observation also validates our motivations to design the *computationally light dynamic gate* based on the *information-rich* motion compensated residuals.

To further investigate the mechanism of the Residual Driven Dynamic Gate, we plot the activation rates w.r.t different activity class in Penn Action in Figure 2. The more dynamic activities like tennis, jumping jacks and bowling requires significantly more accurate features from the decoded frame compared to the static activities like push ups and strum guitars. This observations further verifies the effectiveness of the Residual Driven Dynamic Gate.

2. Implementation Details

We prepare the dataset using the publicly available FFmpeg [2]. Following [5], we adopt the MPEG-4 format to compress the videos and then retrieve the compressed information including Motion Vector and Residuals.

Following [3, 7, 4], We crop the I-frame, P-frame, motion vectors and the residual errors using the provided bounding box for Penn Action dataset. While for Sub-JHMDB, we generate the bounding boxes from the puppet mask following [3]. Each frame from a GOP shares one unique bounding box, which is the mean bounding box of the I-frame from the current GOP and the next GOP. This design ensures that our model could be readily deployed to the real world applications without considering human detection from compressed streams and also further reduces the computation complexity, as the per-frame human detection is no longer needed. The cropped frames are then resized to 256×256 to input to the model. Note that the motion vector stores the offset of the matching blocks between the current P-frame and its previous frame. When cropping is performed, we accordingly modify the offset, which equals to change the reference coordinate from the full frame to the cropped frame.

Random augmentation is adopted during training. Other than the traditional techniques like random flipping, rota-

*Corresponding author.

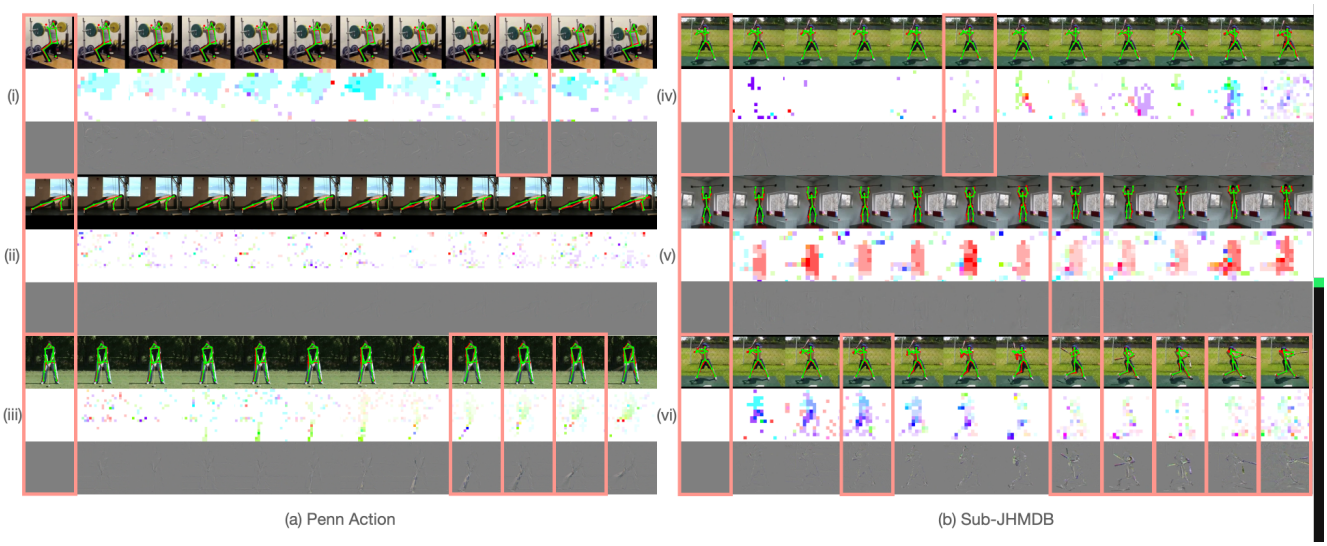

Figure 1. We visualize the decisions of the Dynamic Gate along with the estimated poses for (a) Penn Action and (b) Sub-JHMDB following the same notation as in our main paper. The proposed Dynamic Gate develops a policy to perform feature extractions only for fraction of frames when the GOP contains relatively static sequences while rely more heavily on the accurate features when the pose sequences is dynamic. As depicted in the figure, the frames with faster and more violent motions often come with larger compensation error due to the inaccurate block based motion estimation. Therefore, the decisions of whether to exploit the fully decoded frames for feature extraction could be efficiently determined based on the motion compensated residuals, which explicitly measures the reliability of the motion fields. Furthermore, noticed that in the first half of GOP (v), although no accurate features is provided to our model, our Motion Adaptive Pose Net could still accurately derive the poses for those skipped frames in the pull-up exercise, which also verifies the efficacy of the proposed motion compensated features for fast pose estimations.

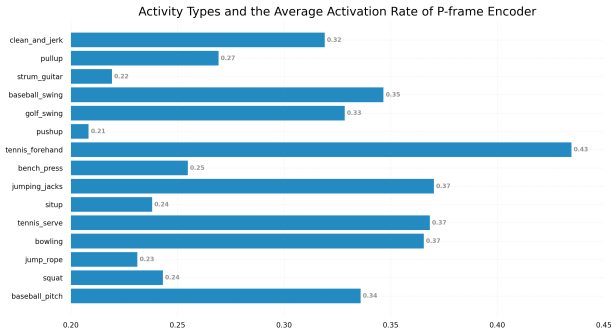

Figure 2. Activity Specific Activation Rate on Penn Action. The most frequently activated activities (tennis forearm) requires twice as many as p frame features compared to the least activated activities (push up and strum guitar).

tion and scaling following [3, 4, 7], we randomly select one frames from the first k frames within the GOP as the I-frame, which is made possible as we retrieve the relative motion vectors and motion compensated residuals instead of back-tracing them to the initial I-frame as [5]. We use $k = 6$ in our experiments. Note that for the motion vectors, other than the traditional augmentation operations, we need to again perform the change of the reference as in cropping.

We employ ResNet18 as our frame encoder in most experiments as it offers comparable accuracy with significantly less computation, as shown by [7] and our comparisons in Table 3 from main paper. We set the hidden dimension of the ConvLSTM d_{hidden} to 64 to balance the computation complexity and the performance of the models based on the validation results. Similarly, we fix the channels of deconvolution layers d_{deconv} to 64. With this design, we allocate most of the computations to the ResNet based encoder following the design philosophy of Simple Baseline [6]. As a comparison, the ResNet18 encoder costs around 4.7 GFLOPs computation, while the decoder only costs 0.72 GFLOPs. Noticed that our pipeline is compatible with any single frame pose estimator. We adopted Simple Baseline [6] structure to maximize the computation savings from the encoding process, which could be dynamically skipped based on the decision of the Residual Driven Dynamic Gate. Following [4, 7], we use the MPII [1] pretrained ResNet encoder to initialize our model.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[2] FFmpeg. Ffmpeg/ffmpeg.

[3] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. Lstm pose ma-

chines. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5207–5215, 2018.

[4] Xuecheng Nie, Yuncheng Li, Linjie Luo, Ning Zhang, and Jiashi Feng. Dynamic kernel distillation for efficient pose estimation in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6942–6950, 2019.

[5] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl. Compressed video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6026–6035, 2018.

[6] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[7] Yuexi Zhang, Yin Wang, Octavia Camps, and Mario Sznaier. Key frame proposal network for efficient pose estimation in videos. In *European Conference on Computer Vision*, pages 609–625. Springer, 2020.