Privacy-Preserving Action Recognition via Motion Difference Quantization

Sudhakar Kumawat and Hajime Nagahara

Osaka University, Japan
{sudhakar,nagahara}@ids.osaka-u.ac.jp

In this supplementary document, first, we provide numerical results corresponding to the plots in the main paper (Section 4.3 and 5.1). Next, we provide visualizations of the output of the BDQ encoder on the three datasets: SBU, KTH, and IPN. Next, we provide reconstruction results output by the 3D UNet network (adversary) on the SBU dataset. Next, we provide an example of a question used in our subjective evaluation. Next, we discuss the feasibility of implementing the BDQ modules using existing hardware. Finally, we provide a section on two studies on the BDQ encoder.

1 Experiments: Numerical Results

	Orig.	**Ideal**	**BDQ**	**Wu**	**Ryoo *et al.*** [2]					
				et al. [4]	$s=2$	$s=4$	$s=8$	$s=16$	$s=32$	$s=64$
Action	90.42%	100%	84.04%	82%	97.93%	98.27%	98.47%	96.27%	92.42%	80.05%
Actor-pair	97.67%	7.69%	34.18%	48%	85.10%	91.48%	84.04%	82.97%	64.89%	43.61%

Table 1: Numerical results on the SBU dataset (main paper - Figure 3). Here, s denotes the down-sampling factor.

	Orig.	**Ideal**	**BDQ**	**Wu**	**Ryoo *et al.*** [2]					
				et al. [4]	$s=2$	$s=4$	$s=8$	$s=16$	$s=32$	$s=64$
Action	92.89%	100%	91.11%	85.89%	91.64%	92.99%	91.22%	91.22%	85.57%	56.21%
Actor-identity	94.34%	4%	7.15%	19.27%	91.82%	92.50%	91.58%	88.86%	82.56%	58.35%

Table 2: Numerical results on the KTH dataset (main paper - Figure 3). Here, s denotes the down-sampling factor.

	Orig.	**Ideal**	**BDQ**	**Wu**	**Ryoo *et al.*** [2]					
				et al. [4]	$s=2$	$s=4$	$s=8$	$s=16$	$s=32$	$s=64$
Action	84%	100%	81%	76%	82.31%	81.76%	79.48%	70.82%	52.96%	31.63%
Gender	90%	50%	59%	65%	80.04%	72.01%	70.08%	64.32%	63.29%	62.70%

Table 3: Numerical results on the IPN hand gesture dataset (main paper - Figure 3). Here, s denotes the down-sampling factor.

2 Ablation Study: Numerical Results

Here, we provide numerical results corresponding to the experiments in Section 5.1 (Figure 4) in the main paper.

	Orig.	**B**	**D**	**Q**	**B+D**	**D+Q**	**B+Q**	**B+D+Q**			
								$\alpha = 0$	$\alpha = 2$	$\alpha = 4$	$\alpha = 6$	$s = 8$
Action	90.42%	89.42%	93.61%	91.48%	88.29%	84.04%	91.48%	89.36%	84.04%	80.85%	75.53%	72.34%
Actor-pair	97.67%	96.00%	90.85%	97.58%	85.10%	59.71%	97.09%	83.12%	34.18%	27.51%	23.40%	22.55%

Table 4: Numerical results on the SBU dataset w.r.t the ablation studies (main paper - Figure 4). Here, α denotes the adversarial weight.

3 Visualizing BDQ Output

Here, we provide visualization results corresponding to the experiments in Section 4.3 in the main paper.

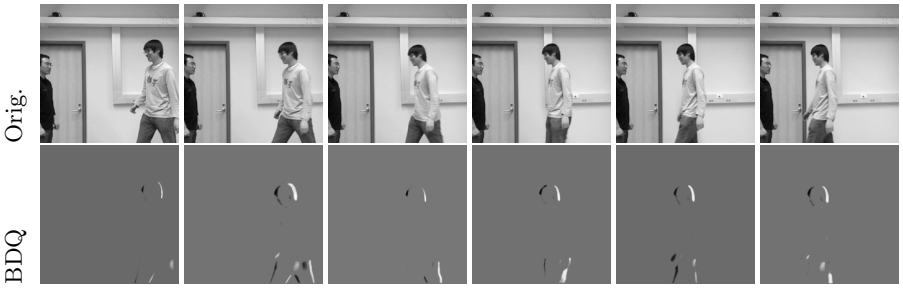

Fig. 1: Example frames of BDQ output on the action video "approaching" from SBU.

Fig. 2: Example frames of BDQ output on the action video "hugging" from SBU.

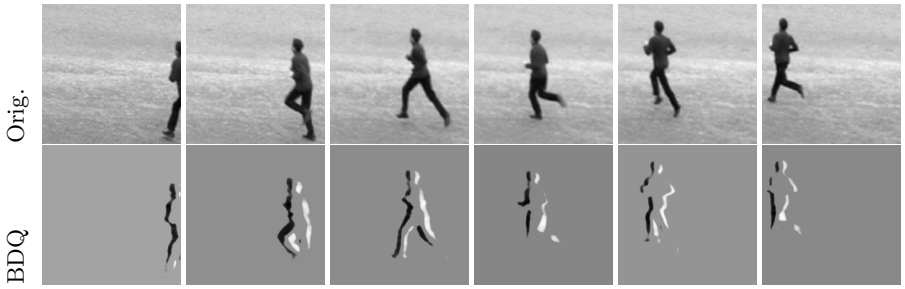

Fig. 3: Example frames of BDQ output on the action video "running" from KTH.

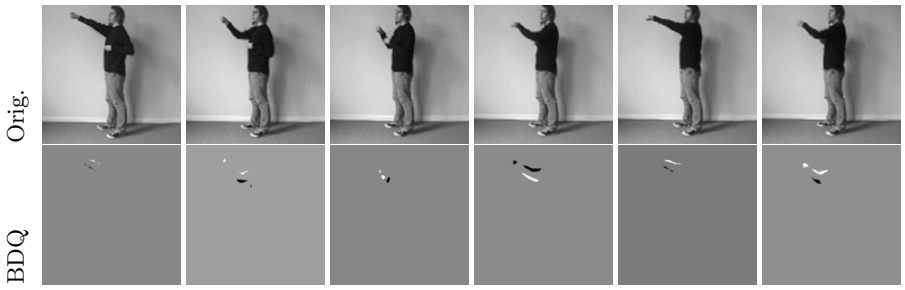

Fig. 4: Example frames of BDQ output on the action video "boxing" from KTH.

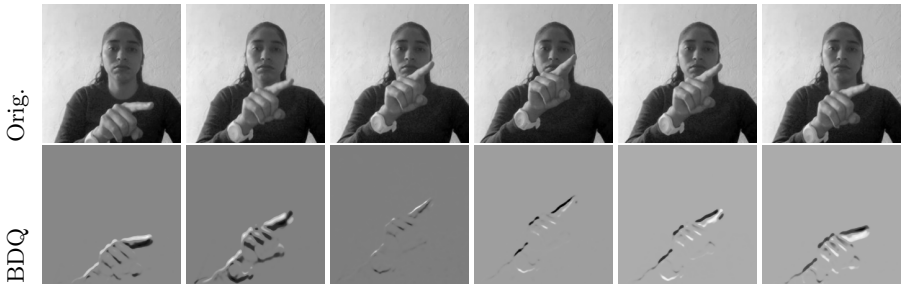

Fig. 5: Example frames of BDQ output on the action video "pointing with one finger" from IPN.

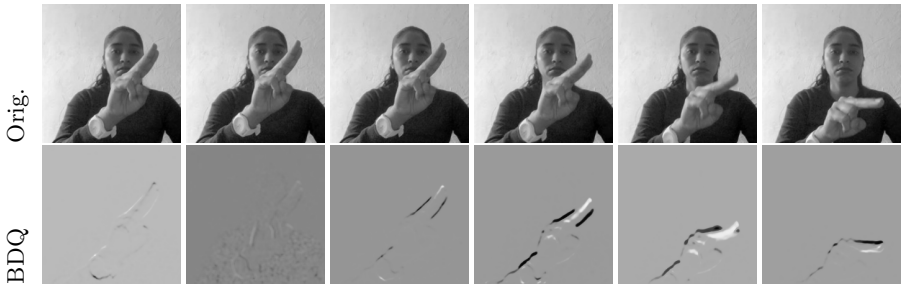

Fig. 6: Example frames of BDQ output on the action video "pointing with two fingers" from IPN.

4 Visualizing Reconstruction Results

Here, we provide visualization results corresponding to the experiments in Section 5.5 in the main paper.

Fig. 7: Example frames of BDQ output, original scene, and reconstruction by 3D UNet adversary on the action video "approaching" from SBU.

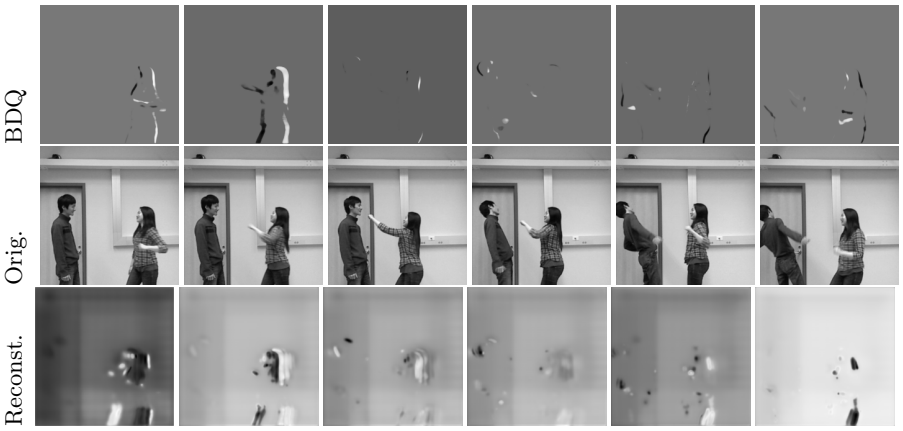

Fig. 8: Example frames of BDQ output, original scene, and reconstruction by 3D UNet adversary on the action video "punching" from SBU.

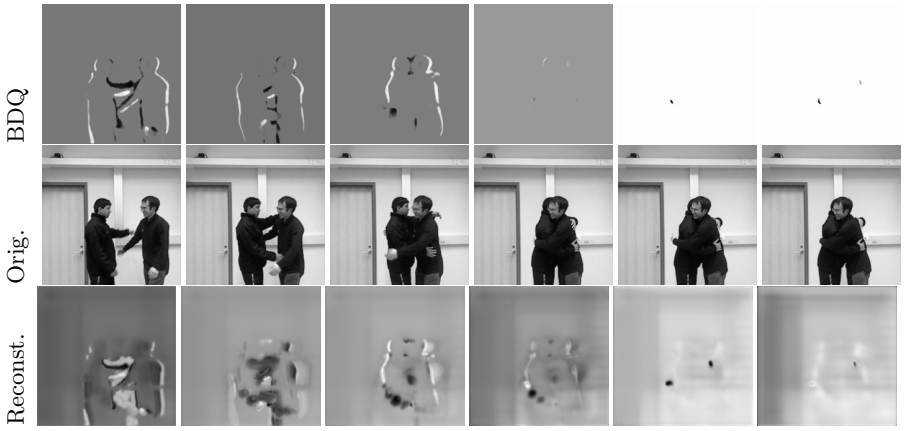

Fig. 9: Example frames of BDQ output, original scene, and reconstruction by 3D UNet adversary on the action video "hugging" from SBU.

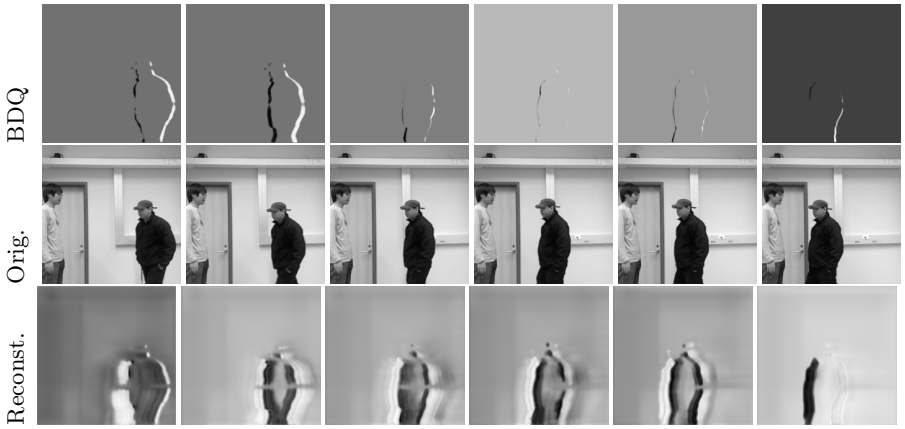

Fig. 10: Example frames of BDQ output, original scene, and reconstruction by 3D UNet adversary on the action video "approaching" from SBU.

5 Subjective Evaluation

In Figure 11, we provide an example of the question from the user study (Section 5.5 main paper).

6 Hardware Implementation Feasibility

This work proposes the BDQ encoder as a software solution for privacy-preserving action recognition. The software solution has its advantages such as cost effective, compatibility with traditional sensors, and easily upgradable over network. However, it is sometimes desirable to implement the encoder into hardware for

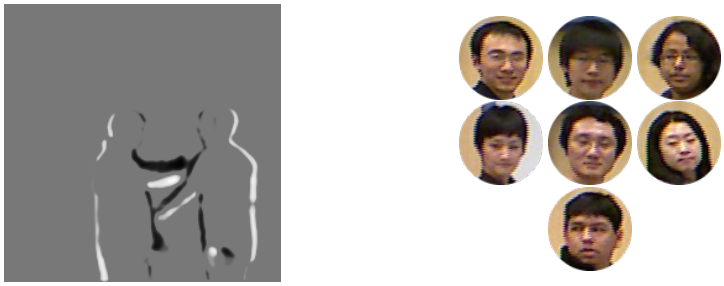

Fig. 11: Left- Example frame of BDQ output used for user study from SBU. Note that, in the original study a .GIF video is played. Right- Seven cropped identities provided as options to select from.

different reasons. Here, we discuss the feasibility of implementing the BDQ encoder using optical and analog computations. As discussed, the BDQ encoder is composed of *Blur*, *Difference*, and *Quantization* modules. From hardware perspective, the *Blur* module can be implemented via optical blur while the *Quantization* module can be implemented by modifying the Analog to Digital circuit (ADC) [3]. The analog implementation of the *Difference* module require a on-pixel frame memory for storing previous frame and a on-chip subtraction circuit. One candidate for the *Difference* module is [1] which propose a CMOS image sensor for motion detection.

7 More Analysis and Studies

7.1 Result of using 3D CNNs for predicting privacy attributes

Table 5 reports the actor-pair (privacy attributes) accuracy on the SBU dataset using various 3D CNNs. It is an interesting finding if how actors move/interact is considered as sensitive information.

Method	**3D ResNet50**	**3D ResNext101**	**3D MobileNetv2**	**3D ShuffleNetv2**
Wu *et al*	50.01 (48.0)	51.47 (49.29)	42.61 (37.12)	42.86 (39.45)
BDQ	38.29 (34.18)	40.42 (33.04)	31.37 (26.45)	34.04 (25.60)

Table 5: Performance results of various 3D CNNs on the actor-pair accuracy. Here (·) denote accuracy of 2D counterpart.

7.2 Sensitivity analysis of the scalar hardness term H.

Table 6 provides this analysis on SBU when 3D ResNet-50 is used for action recognition and 2D ResNet-50 is used for actor-pair recognition. We set $H = 5$ in all our experiments.

	$H = 1$	$H = 5$	$H = 10$	$H = 15$	$H = 20$
Actor-pair	84.39%	34.18%	33.61%	33.54%	33.40%
Action	88.29%	84.04%	81.91%	80.85%	78.72%

Table 6: Performance analysis of the scalar hardness term on the SBU dataset.

References

1. Muramatsu, Y., Kurosawa, S., Furumiya, M., Ohkubo, H., Nakashiba, Y.: A signal-processing cmos image sensor using a simple analog operation. IEEE Journal of Solid-State Circuits **38**(1), 101–106 (2003)
2. Ryoo, M.S., Rothrock, B., Fleming, C., Yang, H.J.: Privacy-preserving human activity recognition from extreme low resolution. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
3. Tan, J., Khan, S.S., Boominathan, V., Byrne, J., Baraniuk, R., Mitra, K., Veeraraghavan, A.: Canopic: pre-digital privacy-enhancing encodings for computer vision. In: 2020 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2020)
4. Wu, Z., Wang, H., Wang, Z., Jin, H., Wang, Z.: Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)