Spatio-temporal Tendency Reasoning for Human Body Pose and Shape Estimation from Videos

Boyang Zhang, Suping Wu*, Hu Cao, Kehua Ma, Pan Li, Lei Lin

* represents corresponding author

School of Information Engineering, Ningxia University, Yinchuan, China

Abstract

- The existing human pose and shape estimation from videos methods are difficult to reconstruct a reasonable human body in an unconstrained environment (extreme illumination, motion blur). Although some approaches attempt to improve performance by adding external data resources, these methods do not take full advantage of the potential information in the underlying data.
- While these methods improve the temporal consistency of human pose estimation in video, they lack spatial understanding and reasoning capabilities, leading to biased predictions.

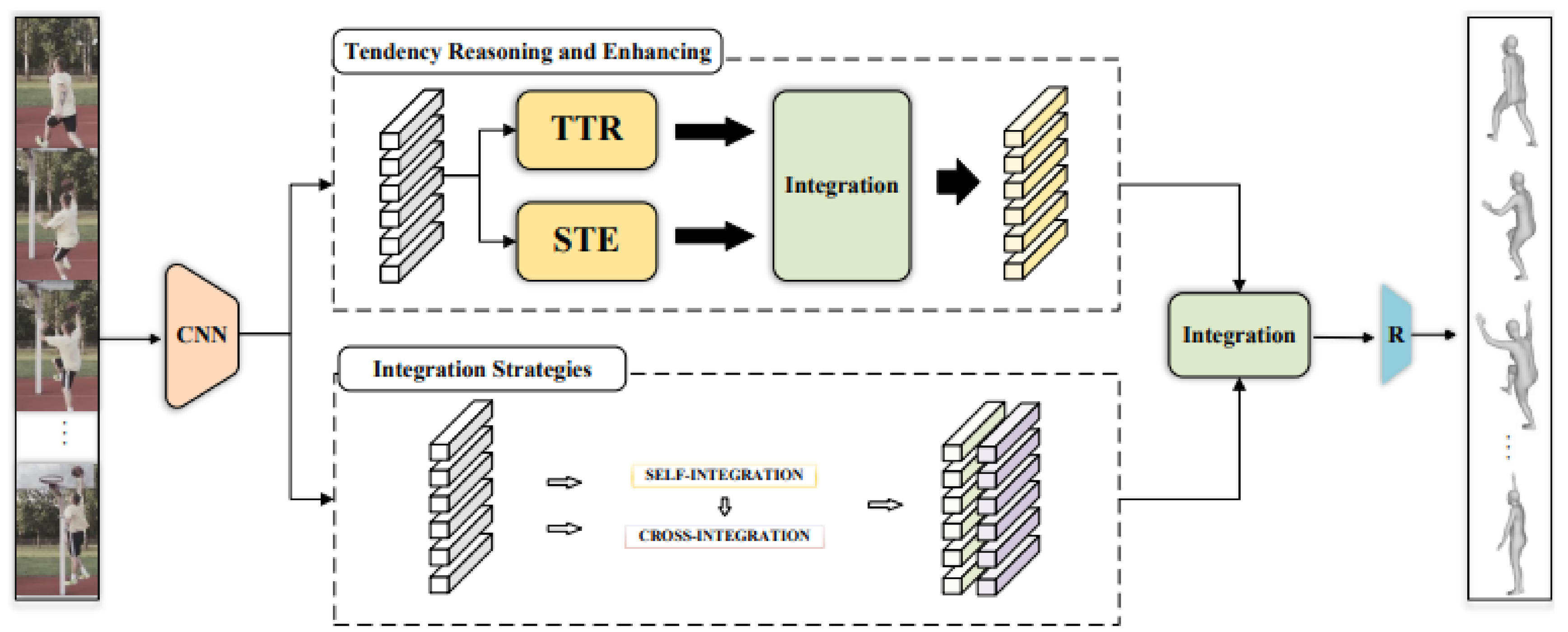
Objective

Our approach aims to alleviate the problem of human reconstruction in unconstrained scenes by reasoning about the spatio-temporal tendency of moving human body.

Main Contribution

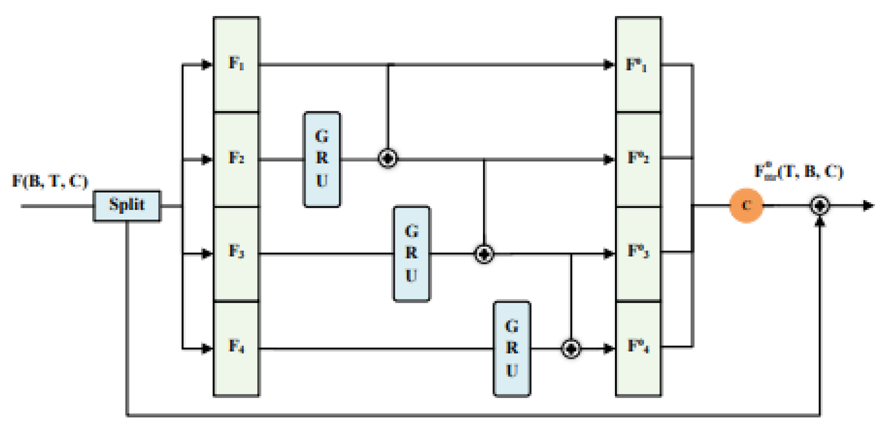
- We propose a spatio-temporal tendency reasoning (STR) for human body pose and shape estimation from videos, which can alleviate the problem of human reconstruction in unconstrained scenes.
- We design a temporal tendency reasoning module and a spatial tendency enhancement module, respectively, to facilitate the effective propagation of motion information over long-distance frames and to stimulate spatially sensitive features. We also propose integration strategies module to enhance the integrated representation of different features.

STR architecture



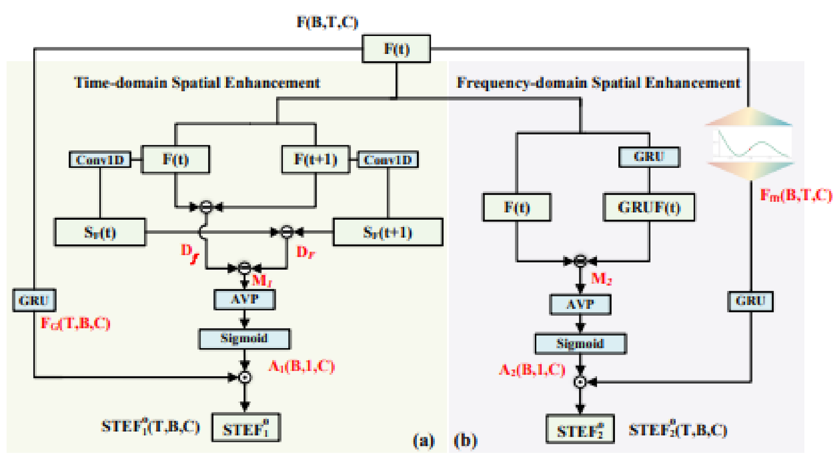
STR consists of two modules, tendency reasoning enhancing module and an integration strategy module. The tendency reasoning enhancing module consists of a temporal tendency reasoning and a spatial tendency enhancing module. The integration strategies consist of a self-integration strategy and a cross-integration strategy.

Temporal Tendency Reasoning



TTR can aggregate temporal tendency across multiple fragments to reason temporal tendency across whole motion sequences. This not only explores long-term dependencies between fragments but also captures information from long-distance frame.

Spatial Tendency Enhancing



STE exploits the temporal differences of adjacent frame-level features to focus on motion features while suppressing irrelevant information in the background. We elaborated on two parts of STE, which are time-domain spatial enhancement and frequency-domain spatial enhancement.

Integration Strategies

Algorithm 1 Integration strategies

Input: All frame features F, number of integration selectable N.
Output: Enhancing features $\hat{F}$

1: /* Splitting F into c parts */
2: Get $F^{c_1}, F^{c_2} = \mathrm{GRU}(\mathrm{SPLIT}(F))$
3: <PHASE 1: SELF-INTEGRATION PHASE>
4: **for** $i < N$ **do**
5: Get $F_i^{c_1} = \mathrm{Integration}(F^{c_1})$
6: Get $F_i^{c_2} = \mathrm{Integration}(F^{c_2})$
7: **end for**
8: <PHASE 2: CROSS-INTEGRATION PHASE>
9: Get $\hat{F}_{SF} = \mathrm{Integration}(F_i^{c_1}, F_i^{c_2})$
10: Get $\hat{F}_{CF} = \mathrm{Integration}(F^{c_1}, F^{c_2})$
11: **return** $\hat{F}_{SF}, \hat{F}_{CF}$

Loss Function

L2 loss was applied to 2D and 3D joint coordinates and SMPL parameters during training.

$$L_{\mathcal{G}} = \omega_{3d}\sum_{t=1}^{T}\|X_t-\hat{X}_t\|_2+\omega_{2d}\sum_{t=1}^{T}\|x_t-\hat{x}_t\|_2+\omega_{shape}\|\beta-\hat{\beta}\|_2+\omega_{pose}\sum_{t=1}^{T}\|\theta_t-\hat{\theta}_t\|_2$$

where X_t stands for 3d joints, x_t for 2d joints, θ and β represent the SMPL parameters. $\omega(\cdot)$ denotes the corresponding loss weights.

Experimental Results

Table 1. Comparisons of our approach with state-of-the-art methods on 3DPW(in-the-wild), MPI-INF-3DHP(outdoor), Human3.6M(indoor) testing set. We denote whether 3DPW is involved in the training process as w/ 3DPW, w/o 3DPW respectively.

	Method	3DPW				MPI-INF-3DHP			Human3.6M		
		MPJPE↓	PA-MPJPE↓	MPVPE↓	Accel↓	MPJPE↓	PA-MPJPE↓	Accel↓	MPJPE↓	PA-MPJPE↓	Accel↓
single image	HMR [[redacted]]	130.0	76.7	-	37.4	124.2	89.8	-	88.0	56.8	-
	GraphCMR [[redacted]]	-	70.2	-	-	-	-	-	-	50.1	-
	SPIN [[redacted]]	96.9	59.2	116.4	29.8	105.2	67.5	-	-	41.1	18.3
	I2L-MeshNet [[redacted]]	93.2	57.7	110.1	30.9	-	-	-	55.7	41.1	13.4
	PyMAF [[redacted]]	92.8	58.9	110.1	-	-	-	-	**57.7**	40.5	-
video	HMMR [[redacted]]	116.5	72.6	139.3	15.2	-	-	-	-	56.9	-
	Sun *et al.* [[redacted]]	-	69.5	-	-	-	-	-	59.1	42.4	-
	VIBE (*w/o 3DPW*)[[redacted]]	93.5	56.5	113.4	27.1	97.7	63.4	29.0	65.9	41.5	18.3
	TCMR (*w/ 3DPW*) [[redacted]]	86.5	52.7	103.2	6.8	97.6	63.5	8.5	73.6	52.0	3.9
	TCMR (*w/o 3DPW*) [[redacted]]	95.0	55.8	111.3 .	**6.7**	96.5	62.8	9.5	62.3	41.1	5.3
	Lee et al. (*w/o 3DPW*) [[redacted]]	92.8	**52.2**	106.1	6.8	**93.5**	**59.4**	9.4	58.4	**38.4**	6.1
	Ours (*w/ 3DPW*)	**85.2**	52.4	**101.2**	6.9	96.3	63.1	8.6	73.3	51.9	**3.6**
	Ours(*w/o 3DPW*)	91.5	55.2	108.7	**6.7**	95.3	61.6	**8.4**	67.8	46.6	**3.6**

Table 2. Effects of the network designs on the performance on the MPI-INF-3DHP dataset.

Model	*PA-MPJPE*↓	*Accel*↓
STR w/o TTR	61.9	8.7
STR w/o STE	62.1	**8.4**
STR w/o STE (time-domin)	61.9	**8.4**
STR w/o STE (frequency-domin)	61.7	8.5
STR w/o Integration strategies(self-)	62.3	9.1
STR w/o Integration strategies(cross-)	62.7	9.0
STR	**61.6**	**8.4**

Figure 1. Qualitative Comparison under extreme illumination.

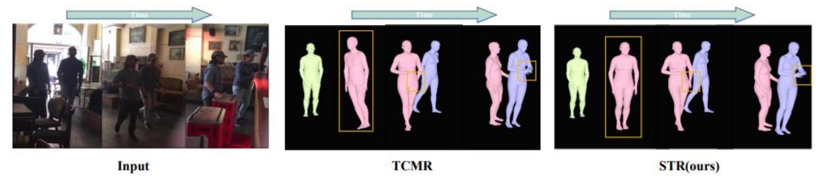


Figure 2. The acceleration errors comparison and Qualitative visualization of STR in the unconstrained scene.