Multi-Sensor Fusion for Human Action Detection and Human Motion Prediction

Thean Chun Koh, Chai Kiat Yeo
Nanyang Technological University
Singapore
KOHT0034, ASCKYEO@ntu.edu.sg

Sunil Sivadas
NCS Pte Ltd
Singapore
sunil.sivadas@ncs.com.sg

***Abstract*—Understanding and predicting human behaviors accurately are essential prerequisites for effective human-robot interaction. Recently, there has been growing interest in multi-sensor fusion for creating robust and dependable robotic platforms, especially in outdoor settings. However, majority of current computer vision models focus on a single modality, such as LiDAR point cloud data or RGB images, and often capture only one person in each scene. This limited approach significantly restricts the effective use of all the available data in robotics. In this study, we propose utilizing multi-sensor fusion to enhance human action detection and motion prediction by incorporating 3D pose and motion information. This approach leverages robust human motion tracking and action detection, addressing issues like inaccurate human localization and matching ambiguity commonly found in single-camera view RGB videos of outdoor multi-person scenes. Our method demonstrates high performance on the publicly available Human-M3 dataset, showcasing the potential of applying multi-sensor multi-task models in real-world robotics scenarios.**

I. Introduction

Precisely comprehending the actions of individuals nearby and forecasting their subsequent movements is pivotal in robotics platforms for ensuring secure and effective interactions between the robot and its environment. To attain this objective, robotic systems must depend on their onboard sensors to gather insightful cues such as human pose to understand human intentions and motions [1], [2].

In recent years, there has been active and ongoing research into 3D human pose estimation algorithms, which utilize either multi-view RGB images [3], [4] or LiDAR point clouds [5] as inputs to predict the 3D human body pose. 3D pose estimation plays a crucial role in various applications, including action recognition [6], human motion prediction [7], augmented reality [8], and robot navigation [9]. However, achieving accurate 3D human pose estimation faces challenges, especially in communities focused on single-sensor inputs, due to several factors. These include individuals often appearing small in images due to their distance from the sensors, leading to difficulties in pose estimation, and frequent occlusions by other individuals or objects, making discernment challenging. Additionally, LiDAR point clouds contain less semantic information, posing challenges in directly recognizing human poses in outdoor environments [10], [11].

To address this issue, the proposed solution involves the integration of multiple sensors to represent 3D poses and motions for both human action detection and motion prediction. Each sensor contributes unique and complementary signals; for instance, cameras capture detailed semantic information, whereas LiDARs offer precise spatial data. Consequently, the fusion of multiple sensors is vital for gaining a comprehensive understanding of human behaviors and accurately forecasting their future movements.

To summarize, the contributions of our paper are:

- We develop a multi-sensor fusion model tailored for detecting human actions and predicting short-term human motion;
- Our model leverages 3D human pose and motion representation to enhance the accuracy of both human action detection and motion prediction;
- Our model is capable of handling scenarios involving multiple individuals in outdoor environments and demonstrates strong performance on the publicly available Human M-3 dataset [10] to highlight the significance and effectiveness of incorporating such multi-modal inputs.

II. Related Works

A. Multi-Sensor Fusion Model

In the domain of computer vision, there has been a notable surge in interest in multi-sensor fusion. The present methodologies can be generally classified into two categories: fusion techniques at the proposal level, fusion techniques at the point level. In proposal-level fusion, exemplified by MV3D [12], object proposals are initially formulated in 3D and then projected onto images to extract region of interest (ROI) features. While proposal-level fusion methods [13], [14] are primarily focused on objects, point-level fusion methods [15], [16] generally superimpose image semantic information onto LiDAR point cloud features, enhancing the point cloud inputs and feature representation. Therefore, these methods are characterized by their dual emphasis on both object and geometric features. Among these techniques, FocalSparseCNN [17] operates at the LiDAR input level, augmenting the data, while DeepFusion [18] operates at the feature level. BEVFusion [19] integrates LiDAR and multi-view RGB images within a shared bird-eye view environment, extracting information from shared geometric and semantic data. Meanwhile, MMVP [10] adopts a voxel-based strategy for multi-modal fusion. Our proposed

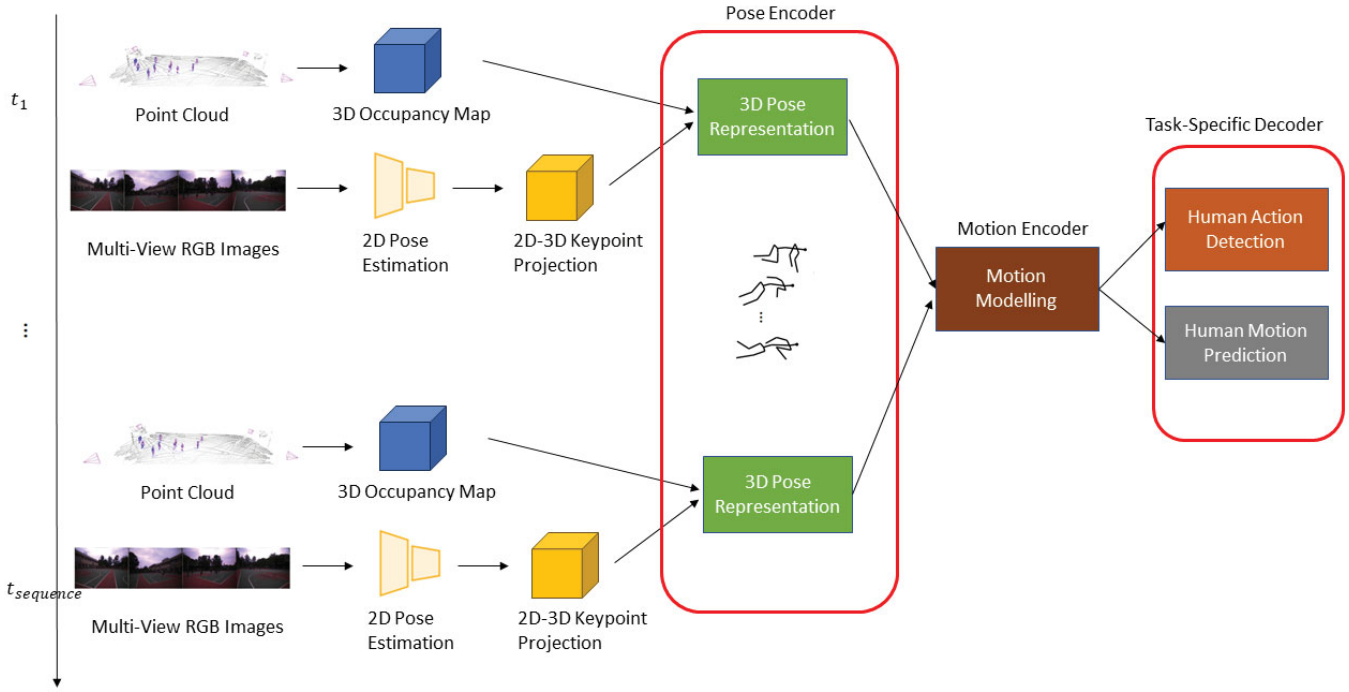

Fig. 1. The overall framework of the proposed multi-sensor fusion to detect human action and predict short-term human motion

model draws inspiration from BEVFusion and MMVP, conducting multi-sensor point-level fusion and extracting human pose and movement within a shared 3D space using a voxel-based approach.

B. Skeleton-based Human Action Detection

The groundbreaking research underscores the intrinsic relationship between human action detection and pose estimation [20], [21]. Previous studies have predominantly utilized LSTM [22] and GCN [23], [24] to capture the spatio-temporal correlation among human joints. Recently, PoseConv3D [25] introduces the idea of stacked 2D keypoint heatmaps, resulting in enhanced outcomes. Alongside traditional supervised action recognition tasks, attention has shifted towards addressing the challenging one-shot and zero-shot human behavior detection problem [26]. For example, [27] utilizes TCN on one-shot detection within therapy scenarios while SL-DML [28] employs deep metric learning on multi-modal input signals. Our approach leverages pretrained motion representations and adapts them to downstream tasks such as human action detection. The pretrain-finetune framework significantly enhances inference performance in scenarios where proper annotation is unavailable.

C. 3D Human Motion Prediction

With access to a few time steps of human motion, we can anticipate the continuation of a person's movement and envision the intricate dynamics of one's future motion. This predictive capability enables us to react and strategize our own actions, particularly beneficial in applications like collision avoidance for robotics [29]. The study of 3D human motion prediction has garnered considerable attention in recent years [30], [31]. For instance, temporal convolution networks [32], [33] have demonstrated promising outcomes in modeling human motion. Although these approaches yield encouraging results, most focus on fixing the pose center and overlook the global body trajectory. Recent studies have begun to address this by jointly predicting human pose and trajectory in the world coordinate system [34], [35]. For example, [36] proposes predicting human motion while considering the constraints of the 3D scene context. CAMP model [37] suggests a two-stage pipeline: first forecasting future contact maps based on past ones and the scene point cloud, then predicting future human poses which rely on the projected contact maps. Similarly, our work forecasts human motion by simultaneously considering 3D poses movements. Furthermore, we extend our predictions beyond individual humans to encompass multi-human motion and interaction.

III. Approach

Fig. 1 shows the overall framework of our proposed multi-sensor fusion. It consists of two encoders and two decoders for multi-sensor fusion and multi-task processing. The two encoders are the pose encoder and the motion encoder, while the two decoders are task-specific decoders: one for human action detection and another for human motion prediction. Before passing the data to the encoders, dataset preprocessing is required to feed the model input.

Initially, the point cloud data undergo voxelization and sampling to generate a 3D occupancy map. Simultaneously, multi-

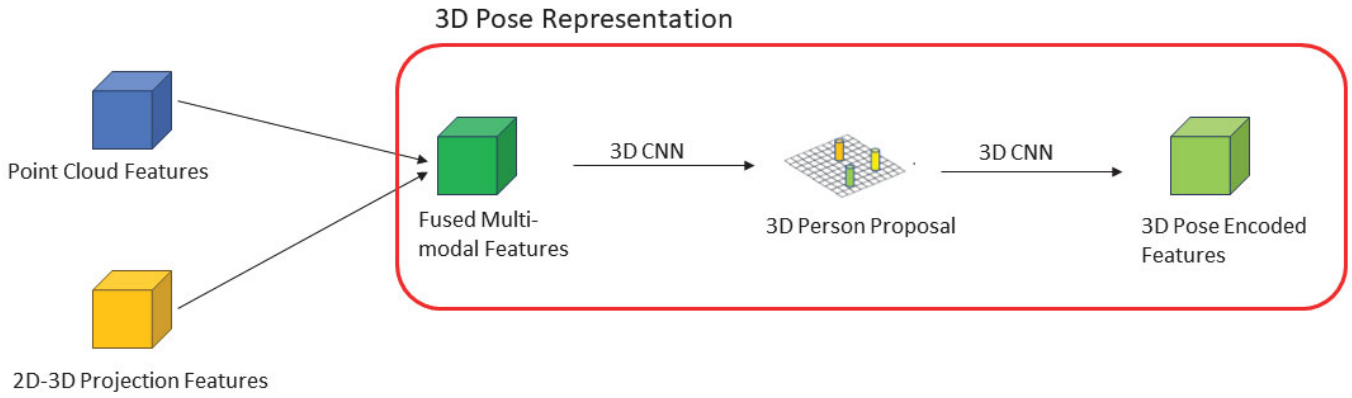

Fig. 2. Architecture of the pose encoder for 3D human pose representation

view RGB images are used to extract 2D human poses through the widely used human pose estimator, RTMPose [38]. These extracted poses are then converted into 2D joint heatmaps, which are subsequently transformed into 3D heatmaps via 2D-3D projection. The resulting 3D occupancy map and 3D joint heatmaps serve as inputs to the pose encoder.

A. Pose Encoder

Point cloud features and 3D joint heatmap are subsequently fused together to form fused multi-modal features within the shared 3D space. These fused features are then passed into a 3D CNN for the purpose of 3D person proposals. Each individual person proposal undergoes a local 3D CNN module, and the outputs from these modules are combined through concatenation to generate 3D pose encoded features, shown in Fig 2.

B. Motion Encoder

As shown in Fig. 3, upon acquiring a predetermined sequence length of 3D pose encoded features, the motion encoder utilizes DSTformer to encode the aggregated features for human motion modeling. DSTformer [26] is a sequence-to-sequence model comprising two branches: one for spatial (SMHSA) and the other for temporal Multi-Head Self-Attention (TMHSA) and MLP. SMHSA captures connections among different joints within a given timestep, whereas TMHSA models the movement of a single joint. By incorporating SMHSA and TMHSA, which respectively capture intra-frame and inter-frame body joint interactions, these fundamental building blocks are combined to fuse the spatial and temporal information in the sequence.

C. Action Detection Decoder

The motion encoded features are projected onto multiple 2D heatmaps through 3D-2D projection. These processed 2D heatmaps are then stacked and fused into a 3D CNN module based on the ResNet 3D backbone to detect multi-person actions within outdoor scenes, shown in Fig. 4. Compared to other fully supervised action detection models, we adapt pre-trained motion representations from the NTU-RGBD dataset and fine-tune the human action detection task using the AVA v2.2 [39] dataset. The pretrain-finetune framework significantly improves inference performance in scenarios where proper annotation is unavailable, such as in the Human-M3 dataset.

D. Motion Prediction Decoder

To forecast human motion while considering 3D pose movements within a scene, we incorporate additional point cloud data alongside the motion encoded features as inputs (see Fig. 5). More specifically, the initial point cloud data within the input sequence length are utilized as supplementary input for the human motion prediction decoder. We employ Point-Voxel CNN (PVCNN) [40] to encode the 3D scene using its DCT feature vectors. PVCNN, specifically designed for processing 3D point clouds, combines voxel-based convolutions and point-based representations, resulting in a memory- and computation-efficient structure for 3D data. In our approach, PVCNN in the decoder is adapted to incorporate motion encoded features as input to predict a residual of the DCT coefficients. This process outputs the predicted short-term movement path, representing global translation. Subsequently, we expand the predictions to include local joint position prediction using a motion prediction module. The motion prediction module in this decoder utilizes a Transformer model, derived from Multi-Range Transformers [31], taking the short path prediction output into consideration. Each individual short path prediction output feeds into the Local-range Transformer Encoder, and the encoded motion features serve as the key and value alongside the query person skeleton for the Transformer Decoder. The resulting output comprises future short-term motion prediction results. The model takes 16 frames as input and predicts the next 32 frames.

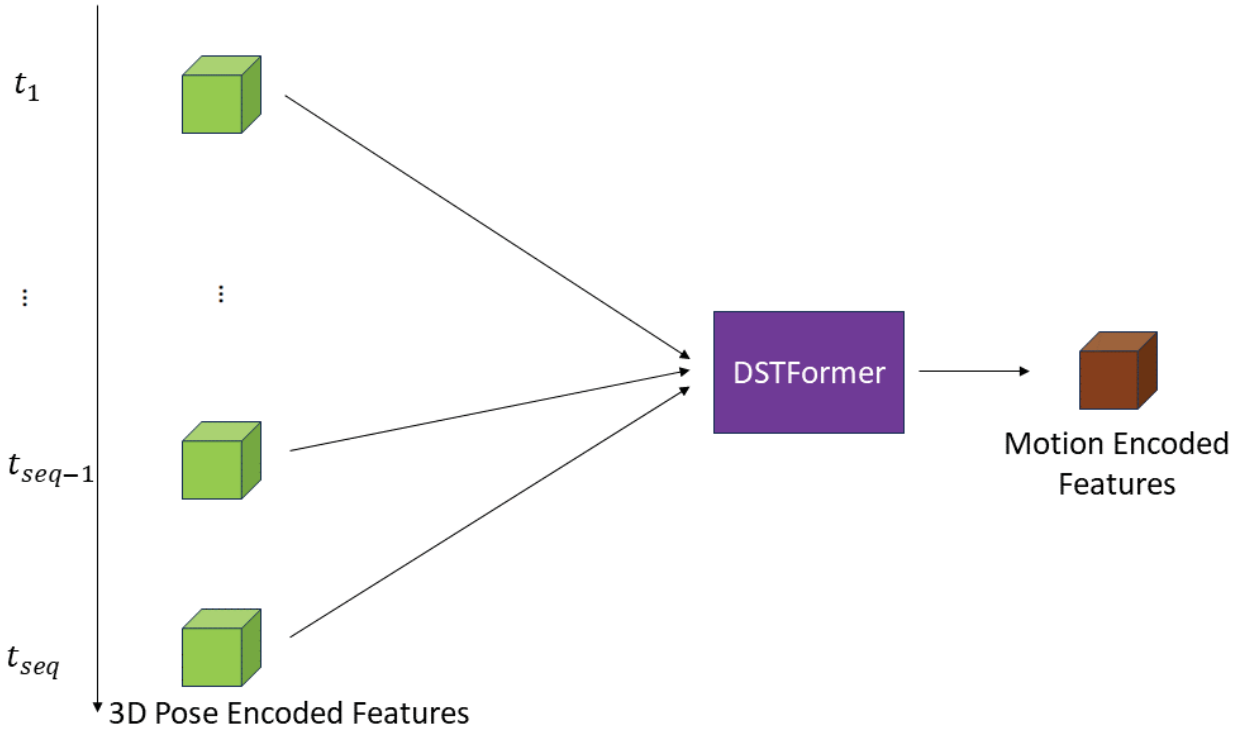

Fig. 3. Architecture of the motion encoder for human motion representation

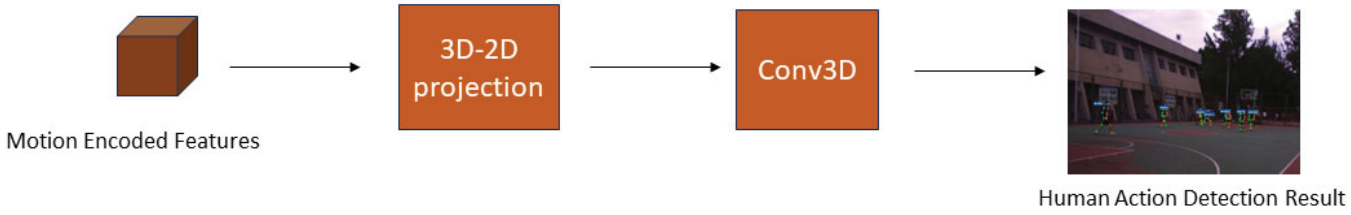

Fig. 4. Architecture of the human action detection decoder

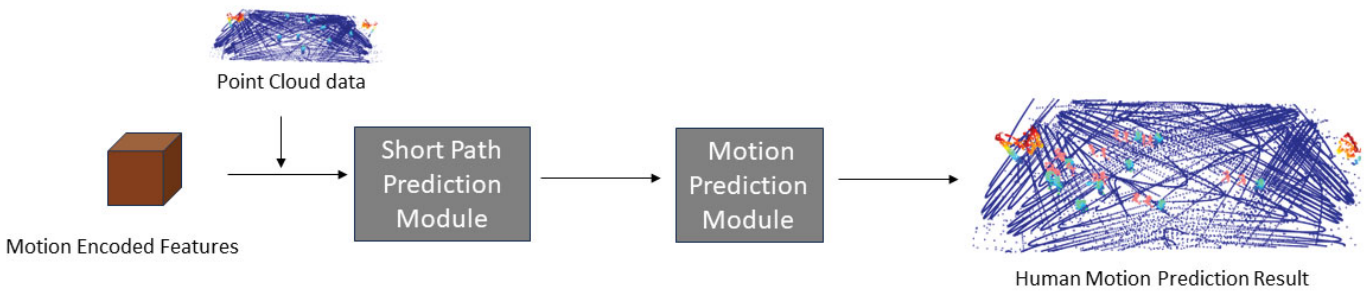

Fig. 5. Architecture of the human motion prediction decoder

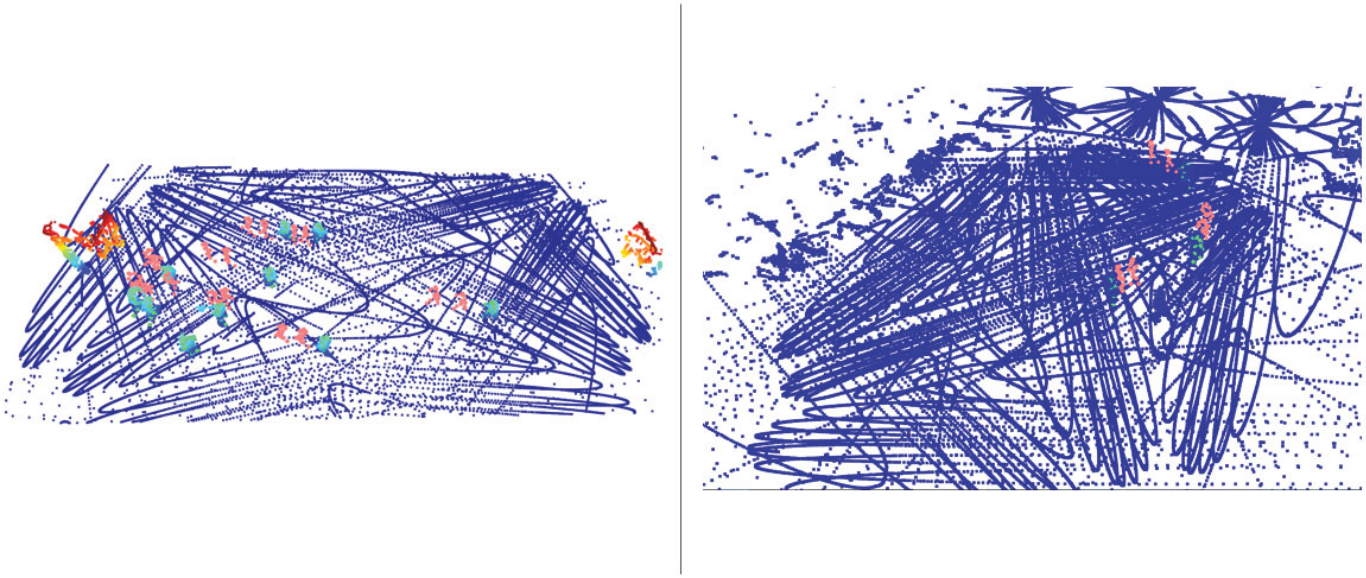

Fig. 6. Visualization result of our model for motion prediction task in point cloud data (basketball court and plaza scene). The green person label represents the current time frame while the orange person label represents the predicted future movement.

Fig. 7. Visualization result of our model for action detection and motion prediction tasks on RGB data. The red person 2D keypoint label represents the predicted future movement that is projected from 3D.

IV. Experiments

A. Dataset

The dataset used for training and validating the proposed model is the Human-M3 dataset. It is an outdoor scene dataset that consists of both multi-view RGB videos and corresponding point clouds. The dataset includes scenarios with multiple persons and provides accurate keypoint labels for each individual. It comprises three main outdoor scenes: a basketball court, a plaza, and an intersection.

B. Implementation Details

In our experiments, both multi-view RGB video frames and point cloud data are sampled with a sequence length of 16 during training and validation. The proposed model undergoes two phases: one for the action detection task and another for the motion prediction task. Despite this division in the training process, the same NVIDIA Tesla V100 GPU is utilized for both tasks' training and validation. Key model configuration settings, include a batch size of 4, an initial learning rate of 0.0001, the use of the Adam optimizer [41], and a weight decay of 0.0001, which remain consistent across both training phases to ensure uniform results with the identical model. The training process for human motion prediction spans 50 epochs, while the training process for human action detection completes in only 10 epochs due to the pretrained-finetune framework.

C. Results and Discussion

TABLE I. PERFORMANCE COMPARISON BETWEEN OUR MODEL AND SOTA METHODS IN TERMS OF ACTION DETECTION ON HUMAN-M3 DATASET.

Model	Accuracy
ST-GCN [23]	34.8
Shift-GCN [24]	35.2
Pose C3D [25]	39.8
Ours	**46.9**

TABLE II. PERFORMANCE COMPARISON BETWEEN OUR MODEL AND SOTA METHODS IN TERMS OF SHORT-TERM MOTION PREDICTION ON HUMAN-M3 DATASET.

Model	Mean Path Error (mm)	Mean Joint Error (mm)
LHMPSC [36]	255.9	130.3
CAMP [37]	245.6	124.1
Ours	**237.2**	**115.4**

Table I presents a performance comparison of our model against state-of-the-art models regarding the human action detection task on the Human-M3 dataset. The state-of-the-art models take single-camera-view RGB videos as input, while our model takes point cloud data and multi-view RGB videos as input. The evaluation metric for this table is the accuracy of human action recognition across frames. The results indicate that our model outperforms the compared models in different scenes, as multi-sensor fusion enhances the performance of motion representation, thus improving human action detection results. Additionally, our pretrained-finetune framework can be easily adapted to new datasets without requiring additional training.

Table II shows a performance comparison of our model against state-of-the-art models concerning human motion prediction on the Human-M3 dataset. This experiment focuses on verifying the performance of the motion prediction decoder. Hence, we adopt a similar input to the state-of-the-art models, which consists of motion encoded features and point cloud data. The evaluation metric is the Mean Per Joint Position Error (MPJPE) [42], which can evaluate both the 3D path and 3D pose prediction. The results demonstrate that our model is capable of achieving great performance in most scenarios and can support multi-person motion predictions, which other compared models are not able to do. Fig. 6 displays the visualization results of the motion prediction on point cloud data, while Fig. 7 visualizes the results of action detection and motion prediction in RGB video frames. There is room for further improvement in the proposed model, especially in terms of complexity. One direction for improvement is to further develop a uniform representation for both pose and motion features.

V. CONCLUSION

In this paper, we propose a multi-sensor fusion framework for detecting human actions and predicting short-term human motion. This architecture focuses on learning 3D poses and motion representation to enhance the performance of both human action detection and motion prediction. It is specifically designed for multiple individuals in outdoor environments and demonstrates strong performance on the Human M-3 dataset. This highlights the significance and effectiveness of incorporating such multi-modal inputs.

ACKNOWLEDGMENT

This study is supported by RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contributions from Singapore Telecommunications Limited (Singtel), through Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU).

REFERENCES

[1] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE transactions on intelligent transportation systems*, vol. 21, no. 3, pp. 900–918, 2019.

[2] J. Li, X. Shi, F. Chen, J. Stroud, Z. Zhang, T. Lan, J. Mao, J. Kang, K. S. Refaat, W. Yang *et al.*, "Pedestrian crossing action recognition and trajectory prediction with 3d human keypoints," *arXiv preprint arXiv:2306.01075*, 2023.

[3] J. Lin and G. H. Lee, "Multi-view multi-person 3d pose estimation with plane sweep stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 886–11 895.

[4] N. D. Reddy, L. Guigues, L. Pishchulin, J. Eledath, and S. G. Narasimhan, "Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 190–15 200.

[5] J. Li, J. Zhang, Z. Wang, S. Shen, C. Wen, Y. Ma, L. Xu, J. Yu, and C. Wang, "Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 502–20 512.

[6] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5137–5146.

[7] J. Yang, Y. Ma, X. Zuo, S. Wang, M. Gong, and L. Cheng, "3d pose estimation and future motion prediction from 2d images," *Pattern Recognition*, vol. 124, p. 108439, 2022.

[8] Y. Su, J. Rambach, N. Minaskan, P. Lesur, A. Pagani, and D. Stricker, "Deep multi-state object pose estimation for augmented reality assembly," in *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 2019, pp. 222–227.

[9] R. Möller, A. Furnari, S. Battiato, A. Härmä, and G. M. Farinella, "A survey on human-aware robot navigation," *Robotics and Autonomous Systems*, vol. 145, p. 103837, 2021.

[10] B. Fan, S. Wang, W. Zheng, J. Feng, and J. Zhou, "Human-m3: A multi-view multi-modal dataset for 3d human pose estimation in outdoor scenes," *arXiv preprint arXiv:2308.00628*, 2023.

[11] Y. Dai, Y. Lin, X. Lin, C. Wen, L. Xu, H. Yi, S. Shen, Y. Ma, and C. Wang, "Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 682–692.

[12] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[13] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1090–1099.

[14] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 172–181.

[15] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.

[16] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 794–11 803.

[17] Y. Chen, Y. Li, X. Zhang, J. Sun, and J. Jia, "Focal sparse convolutional networks for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5428–5437.

[18] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le *et al.*, "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 182–17 191.

[19] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[20] A. Yao, J. Gall, and L. Van Gool, "Coupled action recognition and pose estimation from multiple views," *International journal of computer vision*, vol. 100, pp. 16–37, 2012.

[21] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 915–922.

[22] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[23] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[24] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 183–192.

[25] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2969–2978.

[26] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "Motionbert: A unified perspective on learning human motion representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 085–15 099.

[27] A. Sabater, L. Santos, J. Santos-Victor, A. Bernardino, L. Montesano, and A. C. Murillo, "One-shot action recognition in challenging therapy scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2777–2785.

[28] R. Memmesheimer, N. Theisen, and D. Paulus, "Sl-dml: Signal level deep metric learning for multimodal one-shot action recognition," in *2020 25th International conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 4573–4580.

[29] K. H. Dinh, O. Oguz, G. Huber, V. Gabler, and D. Wollherr, "An approach to integrate human motion prediction into local obstacle avoidance in close human-robot collaboration," in *2015 IEEE International Workshop on Advanced Robotics and its Social Impacts (ARSO)*. IEEE, 2015, pp. 1–6.

[30] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4346–4354.

[31] J. Wang, H. Xu, M. Narasimhan, and X. Wang, "Multi-person 3d motion prediction with multi-range transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6036–6049, 2021.

[32] J. Butepage, M. J. Black, D. Kragic, and H. Kjellstrom, "Deep representation learning for human motion prediction and classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6158–6166.

[33] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 474–489.

[34] Y. Yuan and K. Kitani, "Diverse trajectory forecasting with determinantal point processes," *arXiv preprint arXiv:1907.04967*, 2019.

[35] Y. Yuan, S.-E. Wei, T. Simon, K. Kitani, and J. Saragih, "Simpoe: Simulated character control for 3d human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7159–7169.

[36] Z. Cao, H. Gao, K. Mangalam, Q.-Z. Cai, M. Vo, and J. Malik, "Long-term human motion prediction with scene context," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 387–404.

[37] W. Mao, R. I. Hartley, M. Salzmann *et al.*, "Contact-aware human motion forecasting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7356–7367, 2022.

[38] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen, "Rtmpose: Real-time multi-person pose estimation based on mmpose," *arXiv preprint arXiv:2303.07399*, 2023.

[39] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6047–6056.

[40] Z. Liu, H. Tang, Y. Lin, and S. Han, "Point-voxel cnn for efficient 3d deep learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[42] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.