3D Motion capture technologies for clinical patient monitoring – a short summary

Tamás Karácsony
INESC TEC
FEUP

encontro CIÊNCIA '22

Introduction

3D Human Pose Estimation (HPE) and Motion Capture (MoCap) is a very popular research topic, as it has several applications. However, its application to clinical, patient in-bed monitoring (Fig. 1) is still very challenging, but required for quantitative diagnosis support of epilepsy and sleep monitoring among others. The most promising approaches are the markerless, Deep Learning (DL) based computer vision (CV), 3D MoCap technologies.

Challenges from 24/7 in bed monitoring:

- Continuous occlusions (clinical personnel, blanket)
- At night only low resolution Infrared (IR) B/W and depth videos are available
- Exceptionally irregular, unusual movements during seizures
- Close background
- Markers can not be attached

Fig. 1. Example frame of a person monitored in an Epilepsy Monitoring Unit

Results

Approaches to solve clinical challanges

Occlusions

- **Multiview approaches:** If on one viewpoint a keypoint is occluded on another one it can be still visible. However large memory and computation resource requirements
  - Current top performing approach: TesseTract [2] Fig. 3
    - End-to-end; all feature maps aggregated to a common 4D voxelspace
    - Able to operate in Monocular setting too
  - Spatio-temporal consistency can mitigate short term occlusions (inter-, extrapolation)

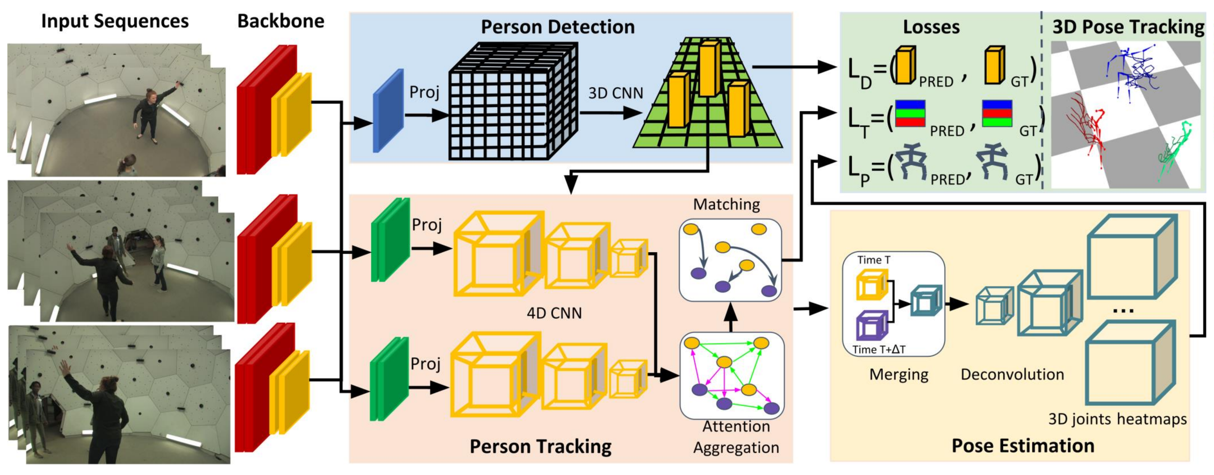

Fig. 3. Full TesseTrack pipeline [2], combines together person detection (3D CNN), tracking (4D CNNs) and pose estimation into one end-to-end network, utilizing 4D voxelspaces.

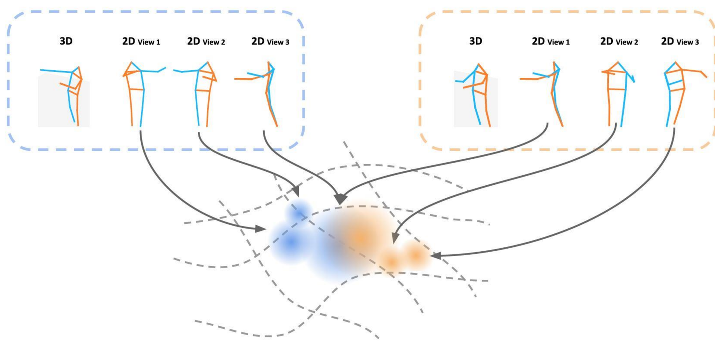

Fig. 4. View invariant embedding of human poses, 2D projections of similar 3D poses are embedded close together, and probabilistic where the same 2D pose projection cover different 3D poses. (Fig. Adapted from [20])

- **Occlusion aware training** - training time augmentation with occlusions
- **Metric learning** – improves view invariance and occlusion robustness
  - Maps close together similar 3D poses and further away different 3D poses in the embedding space (Fig. 4)

Low resolution

- Applying super resolution and image enhancement techniques.
- Train one model for each resolution – impractical
- Resolution aware network with contrastive learning [21]

Video re-colorization

- CNN and GAN based approaches e.g.: VC-GAN [22]
- Temporal consistency is essential to not have flickering of colors

References

[1] C. Ionescu, et al. “Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments,” Tech. Rep., 2014.
[2] N. D. Reddy, , et al. “TesseTrack: End-to-End Learnable Multi-Person Articulated 3D Pose Tracking,” in CVPR2021, 2021, pp. 15 185–15 195.
[3] H. Joo, T. , et al. “Panoptic Studio: A Massively Multiview System for Social Interaction Capture,” IEEE TPAMI, vol. 41, no. 1, pp. 190–204, 1 2019.
[4] T. von Marcard, et al., “Recovering accurate 3D human pose in the wild using IMUs and a moving camera,” in Lecture Notes in Computer Science vol. 11214 LNCS, 2018, pp. 614–631.
5] S. Guan, , et al “Bilevel Online Adaptation for Out-of-Domain Human Mesh Reconstruction,” in Proceedings of the IEEE CVPR, 2021, pp. 10 467–10 476.
[6] D. Mehta, et al. “Monocular 3D human pose estimation in the wild using improved CNN supervision,” in Proceedings - 2017 International Conference on 3D Vision, 3DV 2017, 2018, pp. 506–516.
[7] N. Kolotouros et al., “Learning to reconstruct 3D human pose and shape via model-fitting in the loop,” in Proceedings of the ICCV, vol. 2019-Octob, 2019, pp. 2252–226
[8] L. Sigal, et al “HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion,” International Journal of Computer Vision, vol. 87, no. 1-2, pp. 4–27, 2010
[9] W. Li, et al. “Exploiting Temporal Contexts with Strided Transformer for 3D Human Pose Estimation,” IEEE Transactions on Multimedia, 2022.
[10] M. Trumble, et al. “Total capture: 3D human pose estimation fusing video and inertial sensors,” in British Machine Vision Conference 2017, BMVC 2017,
[11] Z. Zhang, et al. “Fusing wearable IMUs with multi-view images for human pose estimation: A geometric approach,” in Proceedings of the IEEE CVPR, 2020, pp. 2197–2206.
[12] P. Patel, et al. “AGORA: Avatars in Geography Optimized for Regression Analysis,” 2021, pp. 13 463–13 473.
[13] M. Kocabas, et al. “SPEC: Seeing People in the Wild with an Estimated Camera,” ICCV 2021
[14] G. Varol, et al. “Learning from synthetic humans,” in Proceedings of CVPR 2017, vol. 2017 11 2017, pp. 4627–4635.
[15] Z. Wang, et al. “Predicting Camera Viewpoint Improves Cross-Dataset Generalization for 3D Human Pose Estimation,” in Lecture Notes in Computer Science vol. 12536 2020, pp. 523–540.
[16] V. Srivastav, et al., “MVOR: A Multi-view RGB-D Operating Room Dataset for 2D and 3D Human Pose Estimation,” MICCAI-LABELS, 2018.
[17] A. Kadkhodamohammadi and N. Padoy, “A generalizable approach for multi-view 3D human pose regression,” Machine Vision and Applications, vol. 32, no. 1, 2021.
[18] F. Achilles, et al. “Patient MoCap: Human pose estimation under blanket occlusion for hospital monitoring applications,” in Lecture Notes in Computer Science, vol. 9900 LNCS. 2016, pp. 491–499.
[19] G. Pavlakos, et al. “Expressive body capture: 3D hands, face, and body from a single image,” in Proceedings of the IEEE CVPR, vol. 2019, pp. 10 967–10 977.
[20] T. Liu, et al. “View-Invariant, Occlusion-Robust Probabilistic Embedding for Human Pose,” International Journal of Computer Vision, 2021.
[21] Xu, Xiangyu, et al. "3d human shape and pose from a single low-resolution image with self-supervised learning." ECCV. Springer, Cham, 2020.
[22] Y. Zhao, et al. “VCGAN: Video Colorization with Hybrid Generative Adversarial Network,” arXiv preprint arXiv:2104.12357 2021

Related work

Dataset name	Top Model	Note	Data Year
Human3.6M [1]	TesseTrack [2]	Largest base	2014
CMU Panoptic [3]	TesseTrack [2]	10(RGB-D)+480(VGA)+ +30(HD) camera dome	2016-2019
3DPW [4]	DynaBOA [5]	Best in the wild	2018
MPI-INF-3DHP [6]	SPIN [7]	In & outdoor	2018
HumanEva-I [8]	Lifting Transformer [9]	-	2010
Total Capture [10]	GeoFuse [11]	8 camera, 12 IMU	2017
AGORA [12]	SPEC [13]	Synthetic	2021
Surreal [14]	[15]	Synthetic	2017
MVOR [14]	[15]	Synthetic	2017

TABLE I
CURRENT POPULAR DATASETS FOR EVALUATION OF 3D MOCAP

Datasets

- 3D Mocap, clinical MoCap datasets (Table I, II)

Dataset name	Top Model	Note	Data Year
MVOR [16]	[17]	Clinical multivew RGBD	2018
Patient Mocap [18]	[18]	Synthetic Blanket occlusion	2016

TABLE II
CLINICAL DATASETS FOR EVALUATION OF 3D MOCAP

SOTA 3D markerless MoCap

Human body modelling

- Kinematic, planar and volumetric models e.g: SMPL-X (Fig 2.)

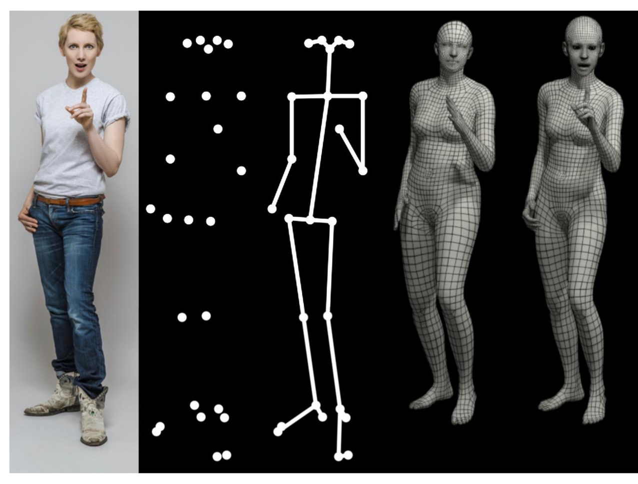

Fig. 2. The SMPL-X model includes, body, hands and face too, with remarkable expressive capabilities. From left to right: Original RGB image, major joints, skeleton, SMPL (female), SMPL-X (female). (Fig. adapted from [19])

Top-down vs bottom-up approaches of 3D MoCap

- **Top-Down:** 1) Detect all individual person 2) Estimate the 3D human poses
  + Take more advantage of body models such as SMPL-X
  – computationally expensive, especially in crowded spaces
- **Bottom-up:** 1) Detect all keypoints 2) Associate keypoints to people
  + Lower computational cost
  - Grouping of joints and occlusions are challanging

RGB Monocular 3D MoCap

- Challangeing due to 3D pose extraction from 2D images can lead to pose ambiguities
- Skeleton only and human mesh recovery approaches with Deep Neural Networks (DNNS)
- Temporo-spatial connections in the DNNs are essential for consistent performance

Depth 3D MoCap

- Resolves depth ambiguity, template based (Fig. 2 - SMPL-X) and template less methods

RGB-D 3D MoCap

- Takes advantage of both color features (RGB) and geometric information (point clouds)

Infrared 3D MoCap

- Virtually non-existent, there are approaches for 2D pose estimation with RGB-IR fusion

Discussion

Challenges

- Several separate solutions to overcome most of the challenges
- Key idea of solutions:
  - Temporo-spatial consistency on every level of the design is essential
  - Learning formulation has to consider guiding the learning process efficiently utilizing:
    - Metric, contrastive learning or triplet loss (control the feature space input of the same pose, these variations include resolution, viewpoint and modality IR/RGB.
    - Occlusion aware training, and end-to-end training to propagate back the error on the whole architecture, improving each sub-task, instead of sequential training.
    - Use prior knowledge such as body models,

Proposed future research direction

- A viable approach can be fusing together different data modalities, here RGB-IR-D, and aim to exploit their separate advantages
- In the end-to-end learning formulation consider metric, contrastive, resolution and occlusion aware training
- Preprocess the IR and RGB videos with super resolutin and re-colorization techniques
- Map all modalities (RGB-D-IR + preprocessed) to a common 4D temporo-spatial voxelspace
- In the 4D voxelspace detect and track the person
- Utilize prior knowledge, such as body model for pose and shape estiamtion, furthermore phisics and kinematics constrains to further refine the 3D MoCap

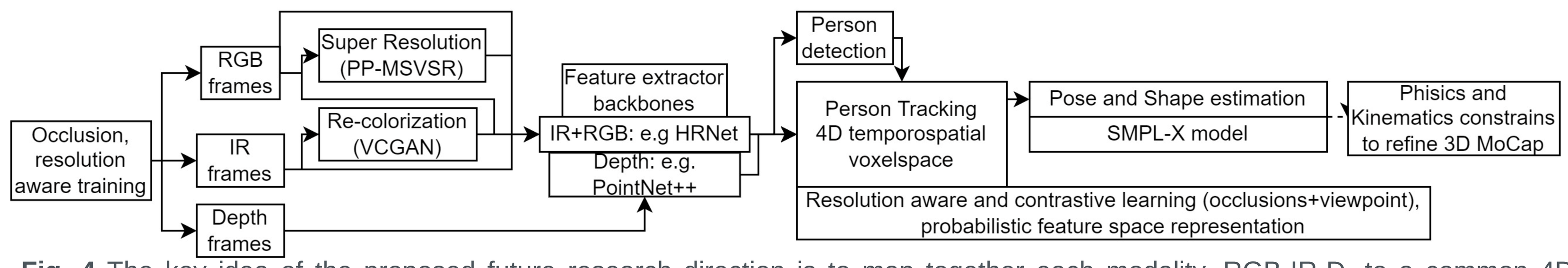

Fig. 4 The key idea of the proposed future research direction is to map together each modality, RGB-IR-D, to a common 4D temporo-spatial volume, extract and improve available features, while constraining the feature space to map the inputs of the same 3D poses and MoCaps close.

Conclusions

In conclusion, markerless 3D Motion capture in clinical environment for patient in-bed monitoring is very challenging, mainly due to heavy occlusions and the requirement of night monitoring. This poster presented the main challenges and existing solutions, furthermore suggested a future research direction.

Acknowledgements

This work was partially funded by Fundação para a Ciência e a Tecnologia under the scope of the CMU Portugal (Ref PRT/BD/152202/2021)

Carnegie Mellon Portugal
FCT Fundação para a Ciência e a Tecnologia
U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO
INESCTEC TECHNOLOGY & SCIENCE