Egocentric Performance Capture: A Review

Shivam Grover, Kshitij Sidana ,Vanita Jain*

Bharati Vidyapeeth's College of Engineering, INDIA
Emails:shivumgrover@gmail; kshitijsidana@gmail.com; vanita.jain@bharatividyapeeth.edu

*Correspondence: vanita.jain@bharatividyapeeth.edu

Abstract

Performance capture of human beings has been used to animate 3D characters for movies and games for several decades now. Traditional performance capture methods require a costly dedicated setup which usually consists of more than one sensor placed at a distance from the subject, hence requiring a large amount of budget and space to accommodate. This lowers its feasibility and portability by a huge amount. Egocentric (first-person/wearable) cameras, however, are attached to the body and hence are mobile. With the rise of acceptance of wearable technology by the general public, wearable cameras have gotten cheaper too. We can make use of their excessive portability in the performance capture domain. However, working with egocentric images is a mammoth task as the views are severely distorted due to the first-person perspective, and the body parts farther from the camera are highly prone to be occluded. In this paper, we review the existing state-of-the-art methods of performance capture using egocentric-based views.

Keyboard: Egocentric Performance; Image Analysis; 3D Animation

1. Introduction

Egocentric vision refers to analyzing images and video feeds from wearable cameras such as those installed on virtual reality headgear and smart glasses. This can be thought of as a first-person view. Motion capture (also referred to as mocap) is the process of capturing the motion of an object or human. This generally deals with recording and understanding the movement of the body and applying it to bigger applications such as animating 3D models and game characters. Egocentric motion capture refers to motion capture from the above-mentioned wearable cameras.

Performance capture and motion capture of the human body is a widely explored area of research. Traditional methods, which have seen tremendous success, usually are costly and require a large amount of space to set up. Marker-based methods [1], [2] require the subject to wear a dedicated costume covered with markings for identifying and tracking. Markerless methods for pose estimation and body reconstruction have seen improvements with the rise of the deep learning era. These however generally require multiple sensors around the subject [3], [4], [5], [6], [7], [8], [9], [10]. These sensors were not attached to the user's body, and they may be static (placed on a tripod) or, in more recent works, dynamic (handheld cameras). This allows the camera to get a more holistic view of the body, which makes it simpler for the artificial intelligence models to understand the motion and the structure of the body. But in areas that are too small to accommodate such an extensive setup or in cases where the subject is

DOI: https://doi.org/10.54216/FPA.020204
Received: March 13, 2020 Revised: April 20, 2020 Accepted: June 19, 2020

constantly mobile, these methods fail since the whole setup cannot be easily accommodated or moved. More recent [11], [12], [13], [14], [15], [16], [17], [18], [19], [20] deep learning-based methods are able to do pose estimation with very few RGB images, but it would still require a camera placed at a distance from the subject. Another case where the performance of such a setup is tested harshly is when there are a large number of objects or people surrounding the

subject. Joo et al. [21] use an array of more than a hundred sensors to capture scenes with dense social interaction. Such setups are incredibly hard to use in a practical way for low-budget and space-constrained applications involving mobile activities.

Using egocentric cameras to perform such an analysis of the subject's body movement seems much more feasible. Since the setup is wearable, it will be portable, it would not depend on the size of the room, and would be unaffected by the environment and people surrounding the subject. While the advantages of an egocentric motion capture system seem good, working with egocentric vision is not such a simple task. The biggest issues are the severely distorted view due to perspective and the high amount of occlusion of the body parts farther away from the cameras. In Fig. 1, one can see the view is at a very oblique angle due to perspective, and the lower body is occluded by the hands. This makes understanding the camera feed much harder compared to the holistic views of the body as used in the traditional methods.

Regardless of the obstacles, some commendable success has been achieved in the field of egocentric vision-based performance and motion capture. These include using outward-facing cameras that are attached to the body for activity recognition [22], [23], [24], [25] and for learning to detect saliency patterns of the users as they interact with the real world [26], [27]. Helmet-mounted cameras have been used [28], [29] to track the head motions and facial features of the subject and the eye gaze using head-worn rigs [30]. Cameras worn on the wrist have been used to track the motion of the fingers and the hand [31]. Foot-worn sensors along with an outside-in depth sensor were used by Zhang et al. [32] to do the full-body pose estimation of the subject. More recent works do a 3D reconstruction of the face [33], [34], the whole body as well as the environment [35].

In this paper, we review the trends in egocentric vision-based performance capture and reconstruction and give an overview of the techniques used. We also look at research that goes a step further and use the motion capture data to reconstruct 3D mesh and textures of the face, body, and environment.

Figure 1: Example views from a head-worn camera that points inside towards the body; oblique angle of the view and occlusion of the lower body can be seen

2. Literature Survey

We segregate the work into different categories based on usage and scope and review each of them one by one.

2.1. Pose Estimation

Pose estimation involves inferring the location of certain key points from images or videos. Each of the key points corresponds to a specific body part or joint.

DOI: https://doi.org/10.54216/FPA.020204
Received: March 13, 2020 Revised: April 20, 2020 Accepted: June 19, 2020

2.1.1.Hand Tracking and pose estimation

Sridhar et al. [36] used only a depth camera to infer the pose of the hand in real-time. They use a randomized decision forest that takes the pixels and classifies them into portions of the hand. This is further optimized by combining the detected part labels with the depth's Gaussian mixture representation. This gives them a pose that corresponds with the depth best. The dataset they use [37] consists of many varying and challenging sequences such as finger counting, waving of the fingers, pinching using fingers, other random motions, etc. To evaluate their model, they take the positions of the 5 fingertips and compute the euclidean error and average it over all the frames, which gives them the lowest average error of 19.6mm. However, they do not consider the interaction of the hands with everyday objects.

Rogez et al. [38] make use of a chest-mounted depth camera to estimate the pose of both the arms and the hands. Their work does consider the interaction of the hands with everyday objects. Instead of relying solely on a local (translation-invariant) scanning window classifier, they use the entire global egocentric view. For the dataset, they render realistic synthetic 3D hand-object data and place them on real 3D backgrounds. This allows them to train their model so that it deals with cluttered scenes.

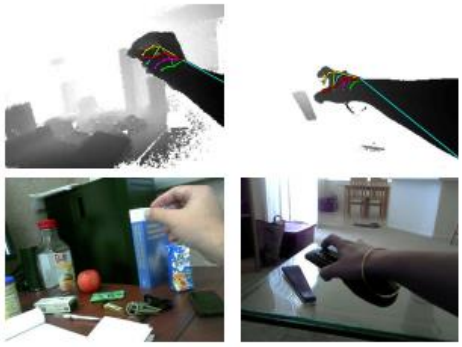

Figure 2: Example results of [38]. The top row shows the results overlayed on the depth image, and the bottom row shows the actual image.

Mueller et al. [39] present a deep learning-based approach to estimate the hand poses while the hand interacts with the objects. They combine two Convolutional Neural Networks (CNNs) for their model. The first CNN estimates the 2D pose of the hand center in the input. This hand position, along with the normalized input depth, is used to generate an image that is normalized and cropped, and sent into another CNN. The second convolutional network now estimates the relative 3D hand joint locations. They are able to achieve this in real-time. They further employ a kinematic pose tracking energy that uses the pose estimates of each frame with a temporal tracking framework. This refines their joint Estimation. To train their network, they curated a dataset, *SynthHands*, which consists of photorealistic data with variations across the pose of the hand, the color of the skin, the shapes and types of objects, the interaction between the hand and the object, and shading details. The dataset is created by rendering a realistic hand model and varying its poses with actual motion capture data of a hand.

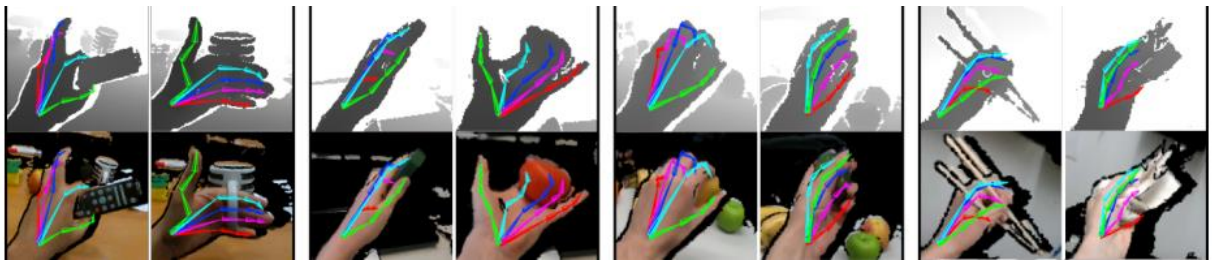

Figure 3: Example results of [39]. The top and middle rows show the predicted pose overlay on top of the input images.

DOI: https://doi.org/10.54216/FPA.020204
Received: March 13, 2020 Revised: April 20, 2020 Accepted: June 19, 2020

2.1.2 Full Body Pose Estimation

In full-body pose estimation, a skeletal model of the whole body is inferred from an image or a sequence of images. The skeletal model may be 2D or 3D, depending upon the scope of the research.

Shiratoriet al. [40] employ 16 body-worn cameras in an inside-out configuration to estimate the full-body 3D pose of the person. They make use of structure-from-motion on individual cameras for a starting point. Then they optimize the reprojection error of the 3D structure and enforce the underlying articulated relationships between cameras and the smoothness of motion temporally.

Jiang and Grauman [41] use a chest-worn camera to infer the body pose. They use the camera to analyze the egomotion and the visible scene instead of directly observing the user. This allows them to infer invisible poses but also leaves them with very limited accuracy. For curating their own dataset, they used a Kinect sensor to obtain the pose while the subject was moving around with the chest-worn camera.

Yonemoto et al. [42] use arm-only RGB-D sensors and indirectly infer torso and arm poses by extrapolating. They further use this data to recognize actions such as opening cover, closing the cover, attaching HDD, etc. They curated their own synthetic dataset for a single human for training.

Rhodin et al. [43] and Xu et al. [44] estimate the full-body pose using head-worn setups in real-time. [43] uses two head-mounted fisheye cameras for stereoscopic vision, whereas [25] uses only a single camera mounted on a baseball cap. For the dataset, [43] recorded 8 people in front of a green screen and used an out-of-the-box motion capture system to get the pose for ground truths, while [44] built a synthetic dataset using the data from [45] and animated the models using the parametrized SMPL model [46] with motions from [47]. [43] reports an average Euclidean error of 70mm for their 3D pose evaluation, while [44] reports the same as 61mm for indoor conditions and 80mm for outdoors.

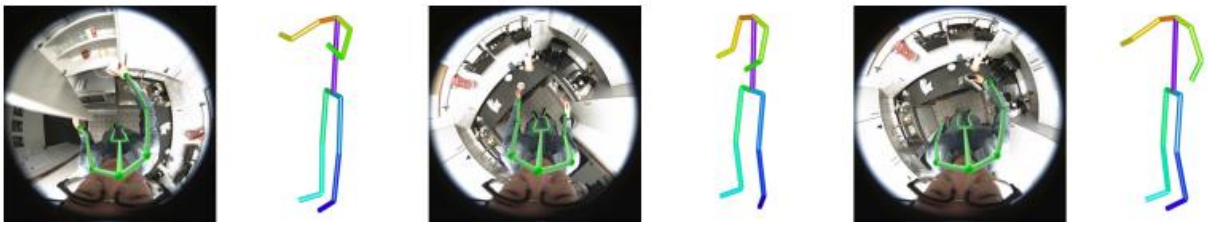

Figure 4: Example results of Mo2Cap2. For each example, the right image is the input to the model overlayed with the predicted pose, and the right image is the 3D representation of the predicted pose.

DOI: https://doi.org/10.54216/FPA.020204
Received: March 13, 2020 Revised: April 20, 2020 Accepted: June 19, 2020

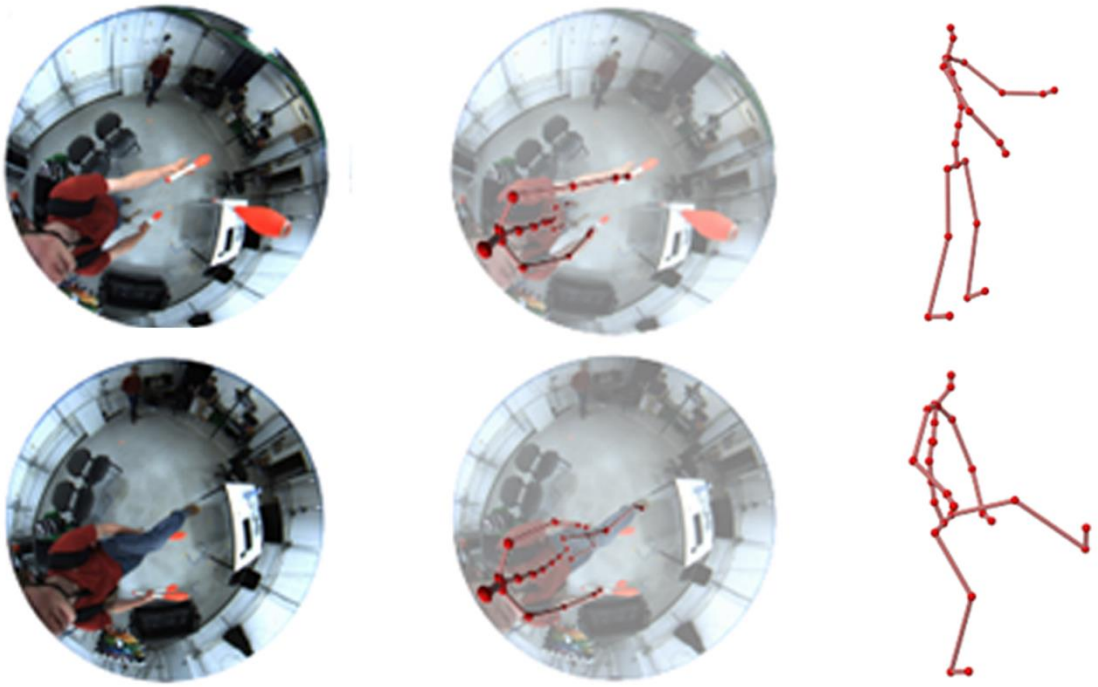

Figure 5: Example results of EgoCap. The right column shows the input to the model. The middle column shows the predicted pose overlayed on the input image, and the right column shows the 3D representation of the predicted pose.

2.2. Gesture and activity recognition

Fathi*et al.* [22] present an approach to understanding activities that one would do on a daily basis, such as preparing dishes from an egocentric point of view. They use a semantic relationship between the activities, actions, and objects to prune the search space which would arise in video interpretation. They use an approach where actions are represented as relations between objects and hands. This allows them to constrain the actions so that they are understood to be followed in a specific order. For example, a person won't start taking out a scoop of ice cream before actually opening the container of ice cream. They get an accuracy of 47.7% in recognizing the activities on a frame-to-frame basis. Kitani*et al.* [23] present an approach to classify actions in first-person sports videos. They use the Dirichlet process to estimate motion codebooks and ego-action categories. The categories include biking, surfing, skiing, horseback riding, snowboarding, etc. They got an F-measure of 0.93 for one of the sequences they tested on, while on the others, they got scores of 0.72 and 0.6.

Ohnishi*et al.* [25] emphasize how handled objects seem small and obstructed in general egocentric images from head-worn and chest-worn cameras. So they employ a camera worn on the wrist to recognize the handled object. They also perform similar experiments with a head-worn camera to compare and show the superior nature of their work. They skip the object detection process as the object is now at a larger scale. They encode LCDs [48] at all the locations in all frames into a single VLAD. This aggregates them into a video representation. They further use CNNs in their network. To collect the dataset, they attached one camera to the head and one to the wrist such that for any given time, there were two frames, one from the head-worn camera and one from the wrist camera. They get an accuracy score of 85.5 for the wrist-mounted camera and an accuracy score of 80 for the head-mounted camera.

Ma*et al.* [24] present a CNN-based architecture that has two streams for classifying activities. The first stream analyzes the information and attributes related to the appearance, and the second stream works with the information

DOI: https://doi.org/10.54216/FPA.020204
Received: March 13, 2020 Revised: April 20, 2020 Accepted: June 19, 2020

related to the motion of the hand. Their architecture is able to learn features that successfully capture object attributes and hand-object configurations. They use the GTEA [49], GTEA gaze [50], and the GTEA gaze+ [50] datasets for their experiments. They get an accuracy of 76.15 for object recognition and an accuracy of 78.33 for action recognition.

Figure 6: Example results of object localization using [24]

Cao *et al.* [51] use a recurrent 3D convolutional neural network for end-to-end learning. They design a spatiotemporal transformer module. This module has recurrent connections between adjacent time slices that can actively create a canonical view from a 3D feature map. To train and evaluate their model, they created a dataset with 83 gestures designed specifically for interactions while wearing wearable devices.

2.2. 3D Reconstruction and Reenactment

3D reconstruction refers to the construction of the 3D mesh and, optionally, the generation of the textures of an object from its image or video.

2.2.1 Facial Reconstruction

This section talks about egocentric-based reconstruction of the whole face and/or inferring the expressions and applying them to the mesh.

Jones*et al.* [52] built a setup that uses an LED-based photometric stereo and augments a head-mounted camera. Their proposed system does not depend on ambient light and allows the generation of per-pixel normals. This allows the performance to be recorded in a 3D dynamic manner. They extend an arm from the headset and attach an LED light to provide artificial controllable light. They estimate the surface normal and recover the surface geometry from it.

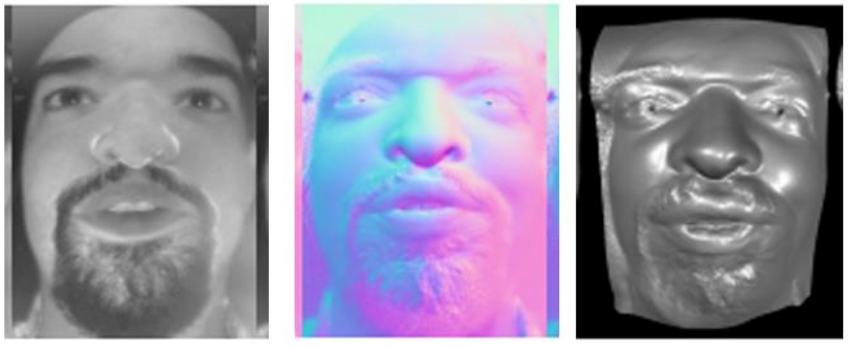

Figure 7: Example results of [52]. The left image is the input from the head-worn camera, the middle image is the estimated normal, and the right image is the recovered surface geometry.

DOI: https://doi.org/10.54216/FPA.020204
Received: March 13, 2020 Revised: April 20, 2020 Accepted: June 19, 2020

Elgharib*et al.* [33] present an approach to achieving reenactment of the facial expressions using an adversarial method. They use a single RGB camera mounted on headwear that points towards the face. They have created a low dimensional latent space, consisting of parameters that are associated with the facial expressions, onto which the input image from the camera is projected. An adversarially trained network is used to transfer the rendered face from the parametric model with the expressions into a realistic video of the person reenacting the actions. They create their own dataset of front face images from their dedicated headgear setup.

Figure 8: Example results of [33]. For both examples, the input to the model is shown on the top left. The ground truth is shown on the bottom left, and the reenacted results are on the right.

One issue that is faced in studying the facial features from head-worn cameras such as in above is that the views are at a very oblique angle, and a major portion of the face is occluded. To handle this, Wei et al. [34] opted for a multiview translation method. They trained a network that allowed them to augment additional views and provide the next reconstruction step with a much more holistic view resulting in much better results.

Figure 9: Example results of [34]. For both examples, the images on the left in gray are the input, and the images on the right are the rendered mesh. The larger grayscale images are the non-augmented input images from the actual headset, and all of the smaller grayscale images are the additional views.

2.2.2 Body and Environment Reconstruction

Cha *et al.* [35] proposed a system that is able to do motion capture and animation of the mesh of the body as well as the face and reconstruct the environment only from the head-worn setup. Their setup consists of multiple cameras which capture both the user and the environment. Their work is divided into three parts: pose reconstruction, face

DOI: https://doi.org/10.54216/FPA.020204
Received: March 13, 2020 Revised: April 20, 2020 Accepted: June 19, 2020

reconstruction, and device tracking and environment reconstruction. Their motion capture works in a similar fashion as [43]. For the facial reconstruction, they use both audio and video data. They detect landmarks in the feed from the cameras to fit a deformable 3D face model. This includes both face shape and expression. They further use a deep neural network that takes as input the audio to make the estimations of the reconstruction more accurate. For training the motion capture module, they record their subject in a data capture setup capable of motion capture for ground truth. Four outwards facing cameras are used to track the user and also reconstruct the environment. Their model is person-specific, which means it is trained only for a single user. They require a prescan for their model to deform using the parameters. For training, they constructed a dedicated data capture setup that is capable of motion capture and recording the subject in it.

3. Conclusion

With a rise in the acceptance of wearable devices in the everyday lives of people, there is a rise in the number of purposes they serve. One such purpose is for recording videos and clicking images. Performance capture from such images is not an easy task, but it has a lot of advantages, including increased portability and feasibility and lower budgets. This review provides a summary of the existing methods of performance capture from a first-person egocentric view. The approaches of the findings and works have been presented and discussed to provide an overview of what has been done so far.

References

[1] A. Woodward, Y. H. Chan, R. Gong, M. Nguyen, T. Gee, P. Delmas, G. Gimel'Farb, and J. A. M. Flores, "A low cost framework for real-time marker based 3-D human expression modeling," *Journal of Applied Research and Technology*, vol. 15, no. 1, pp. 61–77, 2017.

[2] A. Kolahi, M. Hoviattalab, T. Rezaeian, M. Alizadeh, M. Bostan, and H. Mokhtarzadeh, "Design of a marker-based human motion tracking system," *Biomedical Signal Processing and Control*, vol. 2, no. 1, pp. 59–67, 2007.

[3] L. Ballan and G. M. Cortelazzo. "Marker-less motion capture of skinned models in a four camera setup using optical flow and silhouettes". In *3DPVT*. Atlanta, GA, USA, June 2008.

[4] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. "Performance capture from sparse multiview video." In *ACM Transactions on Graphics (TOG)*, vol. 27, p. 98. ACM, 2008.

[5] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. "Marker-less deformable mesh tracking for human shape and motion capture." In Proc. *CVPR*, 2007.

[6] J. Starck and A. Hilton. "Surface capture for performance-based animation."*Computer Graphics and Applications*, 27(3):21–31, 2007.

[7] D. Vlasic, I. Baran, W. Matusik, and J. Popovic. "Articulated mesh animation from multiview silhouettes." In *ACM Transactions on Graphics (TOG)*, vol. 27, p. 97. ACM, 2008.

[8] C. Wu, C. Stoll, L. Valgaerts, and C. Theobalt. "On-set performance capture of multiple actors with a stereo camera."*ACM Transactions on Graphics (TOG)*, 32(6):161, 2013.

[9] G. Ye, Y. Liu, N. Hasler, X. Ji, Q. Dai, and C. Theobalt. "Performance capture of interacting characters with handheld kinects." In Proc. *ECCV*, pp. 828–841. Springer, 2012.

[10] S. L. Colyer, M. Evans, D. P. Cosker, and A. I. T. Salo, "A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System," *Sports Medicine - Open*, vol. 4, no. 1, 2018.

[11] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt, "MonoPerfCap," *ACM Transactions on Graphics*, vol. 37, no. 2, pp. 1–15, 2018.

[12] D. Osokin, "Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose," *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, 2019.

DOI: https://doi.org/10.54216/FPA.020204
Received: March 13, 2020 Revised: April 20, 2020 Accepted: June 19, 2020

[13] Z. Cao, G. H. Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[14] K. Chen, "Sitting Posture Recognition Based on OpenPose," *IOP Conference Series: Materials Science and Engineering*, vol. 677, p. 032057, 2019.

[15] A. P. Yunus, N. C. Shirai, K. Morita, and T. Wakabayashi, "Time Series Human Motion Prediction Using RGB Camera and OpenPose," *International Symposium on Affective Science and Engineering*, vol. ISASE2020, pp. 1–4, 2020.

[16] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization," *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[17] J. Y. Chang, G. Moon, and K. M. Lee, "V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[18] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," *Computer Vision – ECCV 2016 Lecture Notes in Computer Science*, pp. 483–499, 2016.

[19] X. Zhang, D. Zhang, J. Ge, K. Hu, L. Yang, and P. Chen, "Multi-stage Real-time Human Head Pose Estimation," *2019 6th International Conference on Systems and Informatics (ICSAI)*, 2019.

[20] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos, "Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[21] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic Studio: A Massively Multiview System for Social Motion Capture," *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[22] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," *2011 International Conference on Computer Vision*, 2011.

[23] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto, "Fast unsupervised ego-action learning for first-person sports videos," *Cvpr 2011*, 2011.

[24] M. Ma, H. Fan, and K. M. Kitani, "Going Deeper into First-Person Activity Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] K. Ohnishi, A. Kanehira, A. Kanezaki, and T. Harada, "Recognizing Activities of Daily Living with a Wrist-Mounted Camera," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[26] Y.-C. Su and K. Grauman, "Detecting Engagement in Egocentric Video," *Computer Vision – ECCV 2016 Lecture Notes in Computer Science*, pp. 454–471, 2016.

[27] H.S.Park, E. Jain, & Y. Sheikh"3D Social Saliency from Head-mounted Cameras." *NIPS* (2012).

[28] A. Jones, G. Fyffe, X. Yu, W.-C. Ma, J. Busch, R. Ichikari, M. Bolas, and P. Debevec, "Head-Mounted Photometric Stereo for Performance Capture," *2011 Conference for Visual Media Production*, 2011.

[29] J. Wang, Y. Cheng, and R. S. Feris, "Walk and Learn: Facial Attribute Representation Learning from Egocentric Video and Contextual Data," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] Y. Sugano and A. Bulling, "Self-Calibrating Head-Mounted Eye Trackers Using Egocentric Visual Saliency," *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology - UIST 15*, 2015.

[31] D. Kim, O. Hilliges, S. Izadi, A. D. Butler, J. Chen, I. Oikonomidis, and P. Olivier, "Digits," *Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST 12*, 2012.

[32] P. Zhang, K. Siu, J. Zhang, C. K. Liu, and J. Chai, "Leveraging depth cameras and wearable pressure sensors for full-body kinematics and dynamics capture," *ACM Transactions on Graphics*, vol. 33, no. 6, pp. 1–14, 2014.

DOI: https://doi.org/10.54216/FPA.020204
Received: March 13, 2020 Revised: April 20, 2020 Accepted: June 19, 2020

[33] M. Elgharib, R.MallikarjunB.,A.Tewari, H. Kim, W. Liu, H. Seidel, &C.Theobalt(2019). "EgoFace: Egocentric Face Performance Capture and Videorealistic Reenactment." *ArXiv, abs/1905.10822*.

[34] S.-E. Wei, J. M. Saragih, T. Simon, A. W. Harley, S. Lombardi, M. Perdoch, A. Hypes, D. wei Wang, H. Badino, and Y. Sheikh, "Vr facial animation via multiview image translation," *ACM Transactions on Graphics (TOG)*, vol. 38, pp. 1 – 16, 2019.

[35] Y.-W. Cha, T. Price, Z. Wei, X. Lu, N. Rewkowski, R. Chabra, Z. Qin, H. Kim, Z. Su, Y. Liu, A. Ilie, A. State, Z. Xu, J.-M. Frahm, and H. Fuchs, "Towards fully mobile 3d face, body, and environment capture using only head-worn cameras," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, pp. 2993–3004, 2018.

[36] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt, "Fast and robust hand tracking using detection-guided optimization," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[37] S. Sridhar, A. Oulasvirta, and C. Theobalt. Interactive markerless articulated hand motion tracking using RGB and depth data. In *Proc. of ICCV* 2013, pages 2456–2463

[38] G. Rogez, J. S. Supancic, and D. Ramanan. First-person pose recognition using egocentric workspaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4325–4333, 2015.

[39] F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, and C. Theobalt. "Real-time hand tracking under occlusion from an egocentric rgb-d sensor." In *International Conference on Computer Vision (ICCV)*, 2017.

[40] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins. "Motion capture from body-mounted cameras. In *ACM Transactions on Graphics (TOG),*" vol. 30, p. 31. ACM, 2011.

[41] H. Jiang and K. Grauman, "Seeing Invisible Poses: Estimating 3D Body Pose from Egocentric Video," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[42] H. Yonemoto, K. Murasaki, T. Osawa, K. Sudo, J. Shimamura, and Y. Taniguchi. "Egocentric articulated pose tracking for action recognition." In *International Conference on Machine Vision Applications (MVA), May* 2015. doi: 10.1109/MVA.2015.7153142

[43] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, and C. Theobalt. "Egocap: Egocentric marker-less motion capture with two fisheye cameras."*ACM Trans. Graph.*, 35(6):162:1–162:11, November 2016. doi: 10.1145/2980179.2980235

[44] W. Xu, A. Chatterjee, M. Zollhofer, H. Rhodin, P. Fua, H.-P. Seidel, and C. Theobalt, "Mo2Cap2 : Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 5, pp. 2093–2101, 2019.

[45] G. Varol, J. Romero, X. Martin, N. Mahmood, M. Black, I. Laptev, and C. Schmid. "Learning from synthetic humans." In CVPR, 2017

[46] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[47] Carnegie Mellon University Motion Capture Database. http://mocap. cs.cmu.edu/.

[48] Z. Xu, Y. Yang, and A. G. Hauptmann. "A discriminative CNN video representation for event detection." In *CVPR*, 2015.

[49] A. Fathi, X. Ren, and J. M. Rehg. "Learning to recognize objects in egocentric activities." In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference On*, pages 3281–3288. IEEE, 2011.

[50] A. Fathi, Y. Li, and J. M. Rehg. "Learning to recognize daily actions using gaze." In *Computer Vision–ECCV* 2012, pages 314–327. Springer, 2012.

[51] C. Cao, Y. Zhang, Y. Wu, H. Lu, and J. Cheng. "Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. " In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3763–3771, 2017.

[52] A. Jones, G. Fyffe, X. Yu, W.-C. Ma, J. Busch, R. Ichikari, M. Bolas, and P. Debevec. "Head-mounted photometric stereo for performance capture." In *CVMP*, 2011. doi: 10.1109/CVMP.2011.24

DOI: https://doi.org/10.54216/FPA.020204
Received: March 13, 2020 Revised: April 20, 2020 Accepted: June 19, 2020