BODY SHAPE AND CENTRE OF MASS ESTIMATION USING MULTI-VIEW IMAGES

Kentaro Ino[1], Kosuke Takahashi[2], Mariko Isogawa[2], Yoshinori Kusachi[2], Dan Mikami[2], Yuta Sugiura[1], and Hideo Saito[1]

Keio University, Yokohama, Japan[1]
NTT Media Intelligence Laboratories, Yokohama, Japan[2]

This study presents a method for estimating human 3D body shape in action. We propose a method for estimating 3D human body shape motion that uses multiple view images and visual hulls. Related methods necessitated lengthier preparations, such as camera calibration, which would require several tries before actually capturing the image. We solve this issue by combining state-of-the-art computer vision methods to automatically process the required inputs and parameters, so that camera images are the only resource needed for estimation. In our experiments, we applied our method to a video of human subject kicking a soccer ball to left and right side of a goal; we successfully acquired the subject's 3D body shape. In addition, we verified that the application's automatically obtained body shape successfully provides the subject's center of mass.

KEYWORDS: body shape, center of mass, visual hull, image processing

INTRODUCTION: Measuring the movement of athletes is critically important in evaluating their physical performance, with the ultimate aim of enhancing their athletic skills. Currently, various image-processing techniques are being introduced and implemented to provide these measurements. Cao et al. (2017) propose a method of estimating the position of body joints in a single image using deep learning, but such research does not focus on providing a 3D output. One way of extracting 3D output is by using infrared sensors in conjunction with small positional markers (made of infrared-reflective material) attached to the athlete's body; the athlete's movement must then be observed by multiple cameras. Such a method is burdensome and impractical, especially during real sporting events. Our research aimed to develop a practical method of obtaining 3D body-shape movement parameters, as well as additional information, from videos taken with hand-held cameras.

METHODS: Our proposed approach is based on the method of Kaichi et al. (2018) and involves two important elements. First, it enables the use of hand-held cameras by incorporating a camera pose estimation process. Second, its processing is fully automatic. The basic procedure of the proposed method, which requires the preparation of a 3D model of a sports field, is illustrated in Figure 1. When using the system, the user captures a subject with multiple cameras, and the captured images are used for body-shape reconstruction in 3D and for adding kinematic information to the obtained shape.

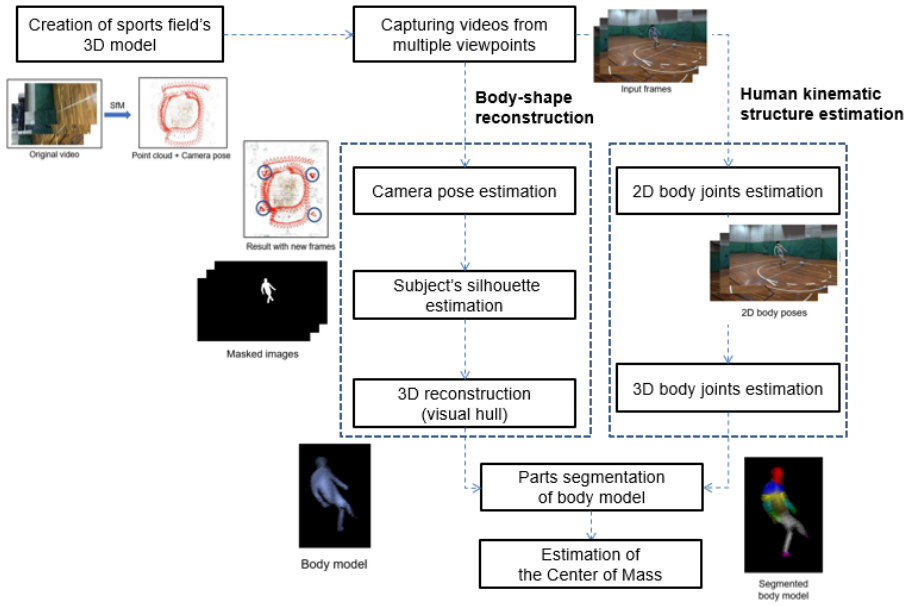

Figure 1: Flow of our method

1. Three-dimensional modelling of field and camera pose estimation: In the computer vision field, before taking videos or images, we must estimate the camera pose. In general, one places a calibration board or similar tool around the target and takes some images to obtain the 2D/3D correspondence, but this is highly impractical and does not work when the target is not stationary. Instead, we use a structure-from-motion (SfM) technique, which estimates the camera pose and generates a 3D point cloud (i.e., a set of points) of the target using multiple-view images. Thus, the video, which is captured by walking around the target before the experiment, provides the only data required beforehand. Using the method of Schönberger et al. (2016), the captured image is then matched with the video frames to find the corresponding points to generate the 2D/3D correspondence in order to estimate the camera pose. At least six correlative points are required for correspondence calculation. The process flow for camera pose estimation is graphically illustrated in Figure 2.

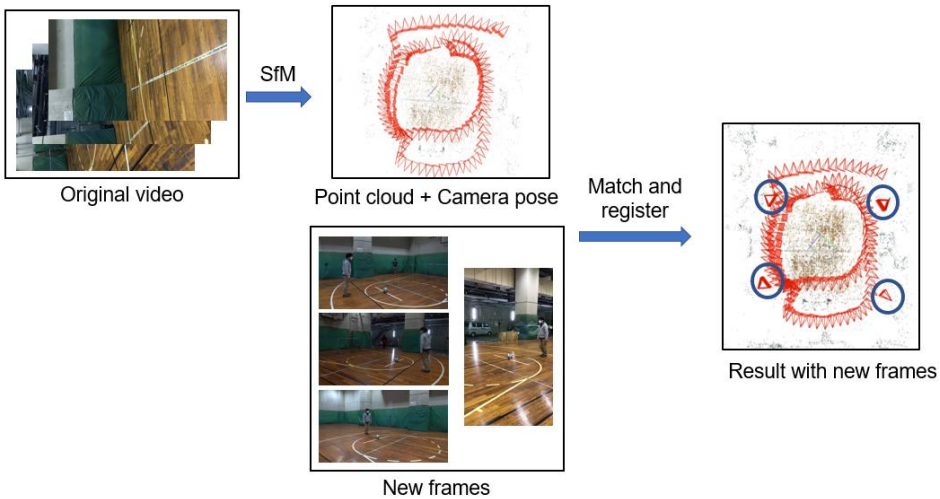

Figure 2: Flow of camera pose estimation

2. Body-shape reconstruction: As illustrated in Figure 1, the body-shape reconstruction requires three steps. Having indicated the first step, camera pose estimation, we now describe the subsequent processes. The 3D reconstruction of the human body is essentially performed using Laurentini's visual-hull method (1994). We extract the 2D silhouette from the input images using Güler et al.'s (2018) method and re-project those silhouettes into a 3D world.

The product set of the silhouettes represents the 3D shape of the body, and V. $V(\ni v_j)$ denotes a set of voxels in which each voxel v_j contains 3D positional information. By reconstructing the 3D shape of the human body, it is possible to reflect any individual's unique figure.
The visual-hull technique uses foreground/background binary images (called silhouette images) and their camera parameters, which indicate the camera poses. The foreground mask is the 2D projection of the corresponding 3D foreground subject. By combining the camera parameters, the silhouette defines a back-projected generalized cone that contains the subject. The intersection of these cones will represent the geometry of the actual 3D object. In the method of Kaichi et al. (2018), the silhouette images are created by manually trimming the body in the images, and the camera parameters are obtained by means of camera calibration. However, both techniques are time-consuming and inefficient. To simplify matters, we produce the silhouette images by a state-of-the-art method that can extract the silhouette automatically, making possible the automation of the system. Our method also allows the use of a handheld rather than a fixed-position camera to obtain the camera parameters.
3. Human kinematic structure estimation: First, we detect the joints of the body. For body segmentation and the removal of noised voxels, we employ Fang's (2017) method for real-time joint detection of an individual's body, hand, and facial keypoints (18 in total), all from a single viewpoint image. By applying a direct linear transform to each 2D keypoint in order to triangulate them, we obtain their 3D positions. Using the segmented model of the body obtained from the body-reconstruction process, we put weight on the segments based on the average of the discrete weights of each part of the body, as described by Leva (1996). The center of mass (CoM) of the body is then calculated as the weighted average position of the voxels.

RESULTS: We applied our method to a video of the test subject kicking a soccer ball to the left and right side of the goal. Four cameras were used in this experiment; one of them was a manually operated handheld smartphone camera, while the other three cameras were positionally fixed. We also estimated the subject's CoM track for each clip, as shown in Figures 3 and 4. Comparing the first frame of each clip (i.e., the blue and black points) reveals that the standing position is different. In addition, each red and green point corresponds to the frame of the subject's impact on the ball, showing that, when kicking to the right side, the subject's body was slightly tilted to the front right as compared to when the subject kicked to the left side.

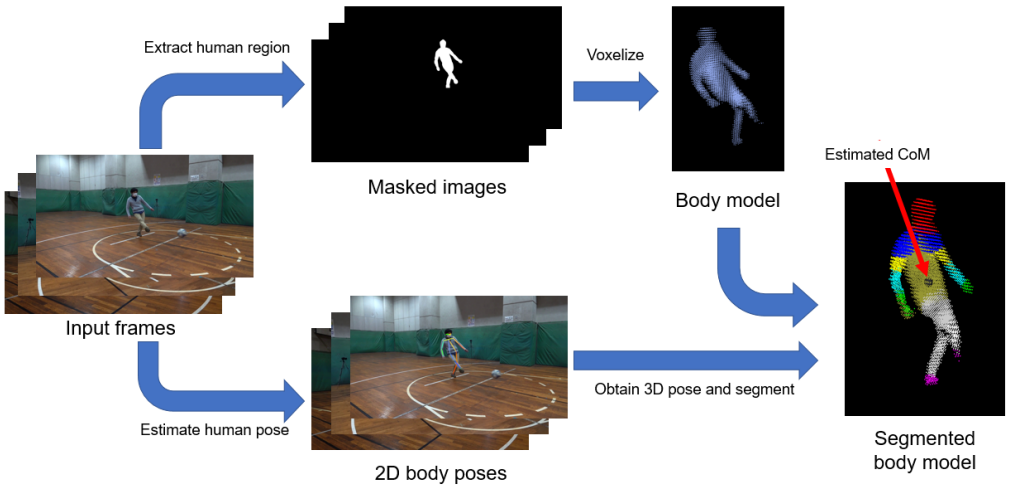

Figure 3: Reconstructed body model and its estimated CoM

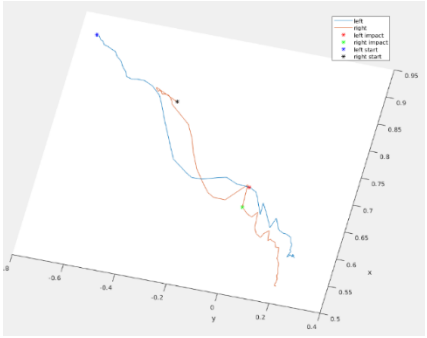

Figure 4: Track of the player's CoM (player moving top left to bottom right)

DISCUSSION: As for the quantitative evaluation, Kaichi et al. (2018) have shown that the error between vertically projected position of estimated CoM and Center of Posture obtained from force plate is about 10 mm. As the process for computing CoM of our method after estimating the projection matrix of each camera is same as Kaichi et al. (2018), we can also expect the similar accuracy of computing CoM in our method.
We have estimated a single person's CoM as an application for body-shape extraction. Though the subject of this experiment was an amateur player, we have shown that there is a significant difference in the manner of kicking. Thus, coaches and sport scientists may discuss how the kicks were different by checking the track of the CoM. As the next practical application, we are considering estimating the CoM of multiple subjects. Although deep-learning-based bone estimation can handle multiple persons in a single view, its challenge is to identify subjects in multiple views. Provided the subject is observed by the camera, any camera can be used for this method. However, since this method estimates the camera pose using 3D modelling of the background, the environment should have enough texture to obtain its 3D model.
For future development, we are now planning to release this software as an open-source product, and we would like to hear the requests of coaches and sport scientists.

CONCLUSION: We proposed a method of extracting the 3D body shape of a person in motion from a multi-view video using novel image-processing techniques. In the experiment, we applied our method to a video of a player kicking a soccer ball to the left and right side of a goal and thereby obtained the body model of the player and the associated track of the player's CoM. In conclusion, we have shown that our method is applicable to the 3D imaging of an athlete's performance in motion.

REFERENCES

Alp Güler, R., Neverova, N., & Kokkinos, I. (2018). Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7297-7306).
Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1302-1310).
De Leva, P. (1996). Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters. *Journal of biomechanics*, *29*(9), 1223-1230.
Fang, H. S., Xie, S., Tai, Y. W., & Lu, C. (2017). Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2334-2343).
Kaichi, T., Mori, S., Saito, H., Takahashi, K., Mikami, D., Isogawa, M., & Kimata, H. (2018). Estimation of Center of Mass for Sports Scene Using Weighted Visual Hull. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1809-1815).
Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. In *IEEE Transactisons on Pattern Analysis and Machine Intelligence,* 16(2): 150-162.
Schonberger, J. L., & Frahm, J. M. (2016). Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4104-4113).

ACKNOWLEDGMENTS: This work was supported by JST AIP-PRISM Grant Number JPMJCR18Y2.