Efficient ConvNet-based Marker-less Motion Capture in General Scenes with a Low Number of Cameras

A. Elhayek, MPI Informatics
E. de Aguiar, MPI Informatics
A. Jain, New York University
J. Tompson, New York University
L. Pishchulin, MPI Informatics
M. Andriluka, Stanford University
C. Bregler, New York University
B. Schiele, MPI Informatics
C. Theobalt, MPI Informatics

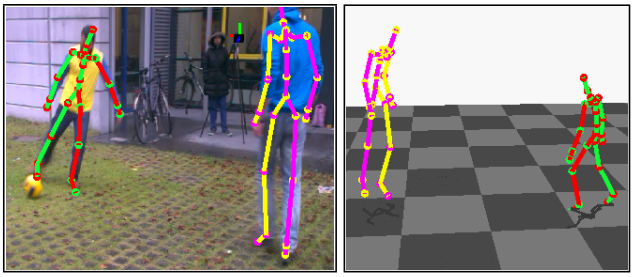

Figure 1. Our ConvNet-based marker-less motion capture algorithm reconstructs joint angles of multiple people performing complex motions in outdoor settings, such as in this scene recorded with only three mobile phones: (left) 3D pose overlaid with one camera view, (right) 3D visualization of captured skeletons.

Optical motion capture methods estimate the articulated joint angles of moving subjects from multi-view video recordings. Motion capture has many applications, for instance in sports, biomedical research, or computer animation. While most commercial systems require markers on the human body, marker-less approaches developed in research work directly on unmodified video streams. Many state-of-the-art marker-less methods rely on a kinematic skeleton model with attached shape proxies, and they track the motion by optimizing an alignment metric between model and images in terms of the joint angles. Formulating and optimizing this usually highly non-convex energy is difficult. Global optimization of the pose is computationally expensive, and thus local methods are often used for efficiency, at the price of risking convergence to a wrong pose. With a sufficiently high number of cameras (≥ 8), however, efficient high accuracy marker-less tracking is feasible with local pose optimizers. Unfortunately, this strategy starts to fail entirely if only $2-3$ cameras are available, even when recording simple scenes inside a studio.

In a separate strand of work, researchers developed learning-based discriminative methods for body part detection in a single image. Detection-based pose estimation can compute joint locations from a low number of images taken under very general conditions. However, accuracy of estimated joint locations is comparably low, mainly due to the uncertainty in the part detections, and pose computation is far from real-time. Also, results on video exhibit notable jitter.

We present a novel method to fuse marker-less skeletal motion tracking with body part detections from a convolutional network (ConvNet) for efficient and accurate marker-less motion capture with few cameras. Through fusion, the individual strengths of either strategy are fruitfully enforced and individual weaknesses compensated. The core contribution is a new way to combine evidence from a ConvNet-based monocular joint detector [2] with a model-based articulated pose estimation framework [1]. This is done by a new weighted sampling from a pose posterior distribution guided by the articulated skeleton model using part detection likelihoods. This yields likely joint positions in the image with reduced positional uncertainty, which are used as additional constraints in a pose optimization energy. The result is one of the first algorithms to capture temporally stable full articulated joint angles from as little as 2-3 cameras, also of multiple actors in front of moving backgrounds.

We tested our algorithm on challenging indoor and outdoor sequences filmed with different video and mobile phone cameras, on which model-based tracking alone fails. The high accuracy of our method is shown through quantitative comparison against ground truth poses. Our approach can also be applied in settings where other approaches for pose estimation with a low number of sensors, that are based on depth cameras or inertial sensors, are hard or impossible to be used, e.g. outdoors.

References

[1] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *ICCV*, 2011. 1

[2] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Join training of a convolutional network and a graphical model for human pose estimation. *NIPS*, 2014. 1