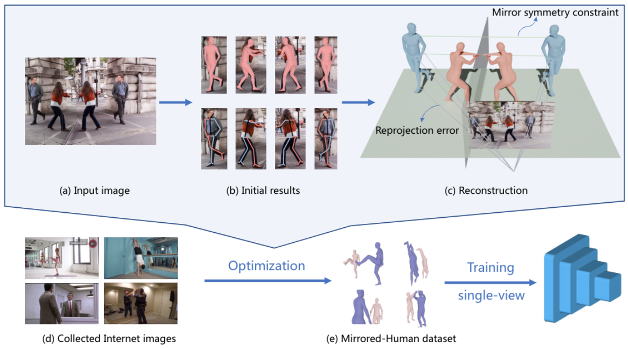
Reconstructing 3D Human Pose by Watching Humans in the Mirror

Fang, Shuai, Dong, Bao, Zhou (CVPR 2021) 04/21/2021 Soma Nonaka



What: Monocular 3D human pose estimation from images that contain a person and his/her image from a mirror.

Why: A mirrored human image provides an additional view that enables us to resolve depth ambiguity and body occlusions.

Mirrors appear in many scenes such as gyms or dancing rooms. In such scenes, we can observe a person with a mirrored image of the person. The mirrored image provides an additional view that can be used to resolve depth ambiguities and body occlusions, which are the innate difficulties of the monocular human pose estimation. This paper is the first to use mirrors for the monocular 3D human pose estimation.

How: The authors derive two constraints from mirror properties; symmetry constraint and normal constraint.

In the proposed pipeline, the authors firstly estimate the 3D pose of a person and the person reflected in the mirror independently. This is done by an off-the-shelf pose estimation model. Next, they refine the two estimated poses under constraints derived from the mirror properties (symmetry constraint and normal constraint).

The symmetry constraint ensures the estimated poses from the original view and the mirrored view have symmetric poses and positions. As Fig. 3 shows, the authors compute lines connecting each pair of body joints between the original and mirrored view (n_i where i is the joint index). These lines should be parallel with each other so that the two estimated poses become symmetric. They therefore introduced a loss term $\Sigma_{(i,j)}||n_i \times n_j||$. In addition, they compute the middle point of the lines (p_i), and introduce a loss term $\Sigma_{(i,j)}||n_i \times (p_i - p_j)||$, which ensures n_i are perpendicular to the mirror plane ($p_i - p_j$). Overall, the loss function is the sum of these two losses (Eq. 4).

The normal constraint ensures the n_i is perpendicular to the mirror normal (n) by introducing a loss function $\Sigma_i||n_i \times n||$. For that, They estimate the mirror normal from two vanishing points. As Fig.4 shows, one vanishing point can be obtained by connecting each pair of 2D keypoints between the original and mirrored person. The other vanishing point can be obtained from mirror edges that they annotated. From these two vanishing points, they estimate the intrinsic parameter of the camera, and use the intrinsic to estimate the mirror normal (Eq. 6). The normal constraint is optional and is omitted when it is difficult to estimate the mirror normal.

When training, the authors used a loss function composed of the symmetric constraints, normal constraints, and a reprojection error between estimated poses and 2D keypoints. The SMPL pose parameters are shared between the original and mirrored person. They use L-BFGS for optimization.

Results: Mirror constraints substantially improve the accuracy of the 3D pose estimation. Generated poses from the proposed model with mirrored human images can be used as pseudo annotations to train another 3D pose estimation model.

As Fig 5 shows, the authors collected videos in which subjects perform various actions in front of a mirror. They annotated 3D keypoints using multiple cameras, and obtained mirror geometry with a calibration board. Correspondences of multiple people are annotated manually.

For comparison, they used SMPLify-X and SPIN. They also created a baseline model by combining the two methods, and used the baseline model to produce an initial estimate that is used in their proposed method.

They trained and evaluated the proposed and existing models with their own dataset. Tab. 1 indicates that considering two constraints (symmetric and normal constraint) greatly improves the accuracy. Even only the symmetric constraint suffices. Qualitative results are apparently good (Fig. 8 and 9). They also examined the accuracy of mirror normal estimation and focal length estimation from the vanishing point, and the error was small (4.1° and 3.6%, respectively).

In addition, they created a dataset by collecting Internet images with mirrors in them and annotating 3D human poses with their method. They use the dataset to train other monocular 3D human pose estimation models. They evaluate the models with 3DPW and Human3.6M datasets (for single-person models) and MuPoTS-3D (for multi-person models). Tab. 3 and 4 show the results. Additional training with their new dataset consistently improves the accuracy.

Thoughts: Since mirrors are often seen in daily scenes, mirror-related techniques will be beneficial.

For example, segmentation of mirror regions will be useful for preventing false detection of people and incorrect matching in SfM.