Three-dimensional pose discrimination in natural images of humans

Hongru Zhu (hzhu38@jh.edu)
Department of Cognitive Science, Johns Hopkins University

Alan Yuille (ayuille1@jh.edu)
Department of Cognitive Science, Johns Hopkins University

Daniel Kersten (kersten@umn.edu)
Department of Psychology, University of Minnesota Twin Cities

Abstract

Perceiving 3D structure in natural images is an immense computational challenge for the visual system. While many previous studies focused on the perception of rigid 3D objects, we applied a novel method on a common set of non-rigid objects—static images of the human body in the natural world. We investigated to what extent human ability to interpret 3D poses in natural images depends on the typicality of the underlying 3D pose and the informativeness of the viewpoint. Using a novel 2AFC pose matching task, we measured how well subjects were able to match a target natural pose image with one of two comparison, synthetic body images from a different viewpoint—one was rendered with the same 3D pose parameters as the target while the other was a distractor rendered with added noises on joint angles. We found that performance for typical poses was measurably better than atypical poses; however, we found no significant difference between informative and less informative viewpoints. Further comparisons of 2D and 3D pose matching models on the same task showed that 3D body knowledge is particularly important when interpreting images of atypical poses. These results suggested that human ability to interpret 3D poses depends on pose typicality but not viewpoint informativeness, and that humans probably use prior knowledge of 3D pose structures.

Keywords: human body, 3d pose, natural images, viewpoint

Introduction

How three-dimensional objects are recognized and represented in the brain has attracted much attention from researchers as a fundamental problem of vision. Recognizing three-dimensional objects is complex, since the two-dimensional projected images of the same 3D object vary considerably as a function of viewpoint, lighting, material and articulation. Recognition is further challenged by the need to recognize image patterns at various levels of abstraction from parts, to individuals, to categories. Many of the early studies focused on the problem of recognizing simple rigid 3D objects given viewpoint variations (Marr & Nishihara, 1978; Biederman, 1987; Ullman, 1989; Tarr & Pinker, 1989; Poggio & Edelman, 1990; Liu, Knill, & Kersten, 1995). However, there has been much less behavioral research on the problem of recognizing non-rigid, articulated 3D objects from natural images, where the range of image variations is considerably larger. This study addresses the problem of recognizing human poses in natural images.

The human body is a stimulus that occurs frequently in daily life and carries a great deal of important information. Our visual system has developed dedicated neural machinery and mechanisms for processing body stimuli (Downing, Jiang, Shuman, & Kanwisher, 2001; Peelen & Downing, 2005). Several behavioral studies on human bodies demonstrated a high degree of sensitivity to properties like gender, mood, identity, etc (Mather & Murdoch, 1994; Ma, Paterson, & Pollick, 2006; Troje & Westhoff, 2006). More recent works studied the representation of body orientation (Lawson, Clifford, & Calder, 2009) as well as body facing directions in the perception of pairs of human bodies (Papeo, Stein, & Soto-Faraco, 2017; Abassi & Papeo, 2020). These previous studies revealed global properties of the human body that support recognition of people's emotions, actions, and social interactions. Yet we still have limited understanding of an important basis for action-related interpretations—the human pose as defined by local body parts and their spatial relationships in three-dimensions.

In this paper, we focused on the perception of 3D poses in natural images of humans, which is particularly challenging due to various joint articulations with different frequency of occurrence, and appearance variations from changes due to occlusion, clothing, lighting, and viewpoint. As natural images may vary in pose typicality and the amount of information views provide to support body part parsing, intuitively we think that human ability to interpret 3D poses in natural images may depend on 3D pose typicality and viewpoint informativeness. To investigate this, we quantified measures of pose typicality to capture the differences between frequent poses (e.g. standing) and less frequent ones (e.g. handstanding) from pose datasets. We also quantified measures of viewpoint informativeness to capture body parsing information primarily based on visibility of joints from projected images. Figure 1 (a) provides example natural images with different 3D pose typicality under different viewpoints. Using a 2AFC pose matching task, we quantitatively measured human performance on interpreting and matching 3D poses in such images. During the task, we asked subjects to match a target natural pose image with one of two comparison, synthetic body images. Observers picked the synthetic image whose 3D pose best matched the target discounting changes in viewpoint about the vertical axis. We expected humans to be better at interpreting typical pose 3D structures, as humans may have more prior knowledge of them. We also expected humans to be better at informative viewpoints where the images contained more useful parsing information.

Our quantitative results showed that human ability to accu-

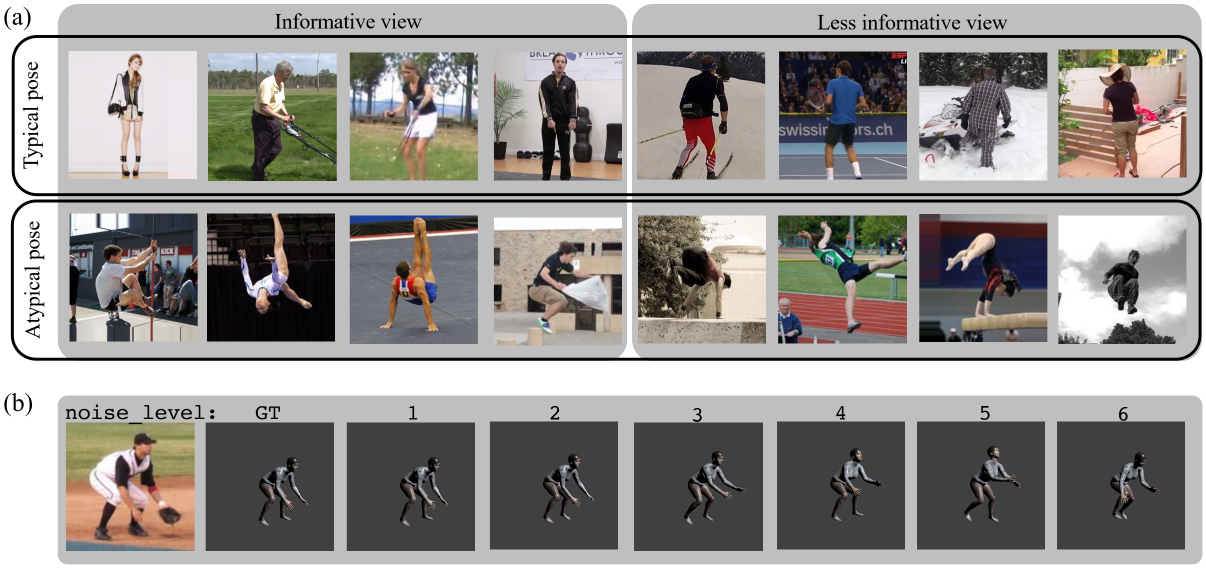


Figure 1: (a) Examples of selected natural pose images. (b) An example image (leftmost) with noise-free synthetic image (second left) as well as noisy synthetic images rendered under different `noise_level` from 1 to 6 with no viewpoint difference.

rately match poses decreased with increasing differences between target and comparison viewpoints. When we grouped trials by typicality of underlying 3D poses and informativeness of viewpoints from natural images, we found that performance for typical poses was measurably better than atypical poses; however, we found no significant difference between informative and less informative viewpoints. These results suggested that the ability to interpret 3D poses in natural images depends on the typicality of the underlying 3D poses but not the informativeness of image viewpoints.

To get further insights on when 3D body information might be more necessary for humans during the task, we built 2D and 3D skeleton-based pose matching models using the same off-the-shelf 3D pose estimator. We tested both models in the pose matching task and compared them under different pose typicality and viewpoint informativeness conditions. Our results suggested that 3D body knowledge is particularly important when interpreting images of atypical poses, and that humans probably use such knowledge of 3D pose structures. Further, our psychophysical results provide useful benchmarks and comparisons to human performance for computational models.

Method

Stimuli Generation

Natural human body images Our stimuli were sampled from the UP-3D Dataset (Lassner et al., 2017), which provides high quality 3D body model fits for single person images from multiple human pose datasets, including MPII (Andriluka, Pishchulin, Gehler, & Schiele, 2014), Fashion-Pose (Dantone, Gall, Leistner, & Van Gool, 2014), LSP and LSP-extended (Johnson & Everingham, 2011).

To ensure that our sampled stimuli were balanced across different poses and viewpoints, we designed and calculated objective scores of pose typicality and viewpoint informativeness for all images before sampling. In general, pose typicality scores captured how similar a pose was to other poses in the dataset. Poses that were dissimilar to the majority were more likely to be infrequent and thus atypical. Viewpoint informativeness scores quantified body parsing information based on visibility of body joints from the image viewpoint.

To obtain *pose typicality scores*, we first defined the distance between two 3D poses. Each pose in UP-3D was annotated with relative 3D rotations of all joints with respect to their parents in the kinematic tree, and thus pose i with K joints was represented by a list of unit quaternions $[q_{i1}, q_{i2}, ...q_{iK}]$ representing these 3D rotations. Distance between two poses i and j was defined as

$$D(i,j) = \frac{\sum_{k=1}^{K} dis(q_{ik}, q_{jk})}{K} \quad (1)$$

where $dis(q_{ik}, q_{jk}) = \arccos(|q_{ik} \cdot q_{jk}|)$ defines the distance between two unit quaternions q_{ik} and q_{jk} (Huynh, 2009).

Pose typicality score was subsequently defined as the average distance from the current pose to all other poses in UP-3D. A pose would be deemed more atypical if it was on average more distant to the rest of the UP-3D poses.

For *viewpoint informativeness scores*, we first calculated f_{ik}—the fraction of pixels for each joint k with respect to pixels for the whole body of pose i. We then standardized these fractions for different joints across all images in UP-3D by

$$\text{z-score}(f_{ik}) = \frac{f_{ik} - \mu_k}{\sigma_k} \quad (2)$$

where μ_k and σ_k are mean and standard deviation of f_{ik} for joint k across all poses in UP-3D. *Viewpoint informativeness scores* were defined as the average z-score from K joints:

$$\frac{\sum_{k=1}^{K} \text{sigmoid}(\text{z-score}(f_{ik}))}{K} \tag{3}$$

where sigmoid was to ensure scores fall in $(0,1)$. A pose would be considered more informative if more joints were visible with larger areas (i.e., more pixels) in the images.

With these objective scores for pose typicality and viewpoint informativeness, we sampled 400 stimuli from UP-3D, with 100 from each of the four categories defined by this 2×2 combination—(typical pose, atypical pose) × (informative view, less informative view). Then we split these 400 stimuli into two groups of 200 stimuli ($Group_1$ and $Group_2$), and tested them on two groups of subjects separately. Figure 1 (a) shows examples of natural pose images selected for use in our experiments.

Synthetic human body images With 3D joint rotation parameters for each natural pose, we used Blender 2.79 to make 2D projections of 3D synthetic humans under different poses.

For each natural pose stimulus, we rendered both noise-free and noisy synthetic humans with constant, predetermined clothing and lighting. For noise-free versions, we posed synthetic humans using original 3D joint rotations from UP-3D. For noisy versions that were used as distractors under a given noise_level, we added $(2\times\texttt{noise\_level}-1)\pi/128$ with a random sign $\{+,-\}$ to the rotation angles in axis-angle representations for all body joints, and we posed synthetic humans using relative 3D joint rotations with added noises. Noise levels were determined by pilot experiments and were set from 1 to 5 ($Group_2$ used an additional $\texttt{noise\_level} = 6$ for control trials). Figure 1 (b) shows an example with noisy synthetic distractors generated under different noise_level.

Once we get all pairs of synthetic humans with and without added joint rotation noises under each noise_level, we rotated both noise-free and noisy synthetic humans horizontally by r degree(s) together so that viewpoints were changed from the original viewpoint in natural images. We uniformly sampled r from $[0°, 15°, 30°, 45°, 60°, 75°, 90°]$.

Psychophysics

Participants Two groups of Amazon Mechanical Turk workers based in US participated ($n_1 = 35$, $n_2 = 42$). Data from subjects with control trial accuracy below threshold ($T_1 = 0.55$, $T_2 = 0.7$) were excluded. Threshold T_2 was higher because control trials in $Group_2$ used more distinct synthetic distractors rendered with no viewpoint rotations at an additional $\texttt{noise\_level} = 6$, whereas control trials in $Group_1$ used synthetic distractors rendered with no viewpoint rotations at $\texttt{noise\_level} = 5$. All reported results were based on the data from the remaining participants ($n_1 = 28$, $n_2 = 33$).

Procedure Each subject went through 200 trials with natural pose images entirely from one of the two groups. Each image was shown exactly once across the experiment.

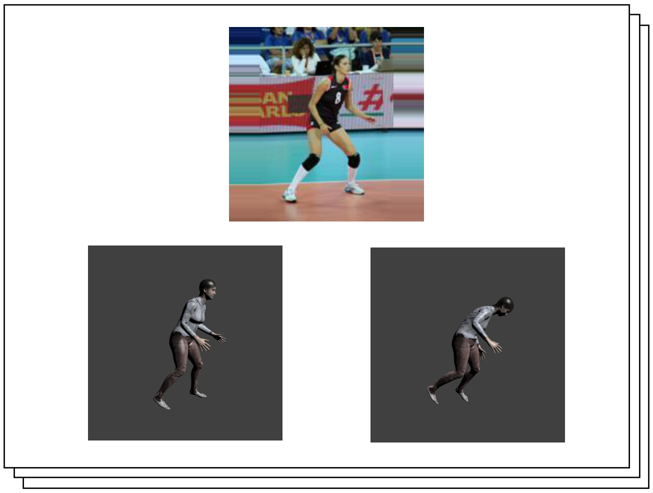
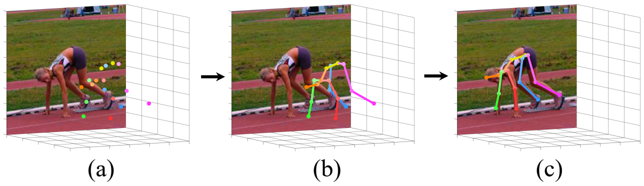


Figure 2: An example experiment trial where the noisy distractor pose was rendered under $\texttt{noise\_level} = 5$. Both noise-free (bottom left) and noisy distractor pose (bottom right) had a 30° viewpoint difference from the target (top).



(a) (b) (c)

Figure 3: An example for 3D and 2D skeleton-based model. (a) Predicted 3D joints. (b) 3D skeleton. (c) Projected 2D skeleton (discarding depth coordinate).

Each subject was given three practice trials before they started the experiment. In each trial, they performed a 2AFC pose matching task on a screen with three images. On the top was the target natural pose image. On the bottom were two comparison, synthetic body images—one was rendered with the same 3D pose parameters as the target while the other was a distractor rendered with added noises on joint angles under a random noise_level. Both synthetic human bodies were rotated by the same random degree(s) before rendered into 2D image projections. These two synthetic body images were randomly placed to the left or right of the screen in each trial. Figure 2 shows an example trial where $\texttt{noise\_level} = 5$ and viewpoint difference was 30°. Subjects were asked to pick the synthetic image whose 3D pose best matched the target discounting changes in viewpoint about the vertical axis. Subjects were given unlimited time to respond to each trial while the three images were constantly shown on the screen. Subjects can choose to take a break after finishing 100 trials.

Pose matching model

3D joint coordinate estimation We used an off-the-shelf monocular 3D human pose estimation model developed by Moon, Chang, and Lee (2019) to estimate 3D body joints from a single image. This model achieved comparable results with the state-of-the-art on 3D single-person pose estimation. In our task, we loaded the model with pre-trained weights on MuCo-3DHP (Mehta et al., 2018) and MSCOCO (Lin et al., 2014), and applied it on both natural and synthetic body images to get predicted joints and 3D skeletons (Figure 3 (a-b)).

Skeleton-based pose matching We transformed predicted 3D skeletons into view-invariant representations for pose matching. Specifically, this representation is a list of joint angles from a skeleton. A joint angle is the angle between the two bones on either side of a joint. For example, given predicted joint positions of left shoulder x_{ls}, left elbow x_{le}, and left wrist $x_{lw} \in \mathbf{R}^3$, the joint angle at the left elbow is

$$a_{le} = \arccos \frac{(x_{ls} - x_{le}) \cdot (x_{lw} - x_{le})}{|x_{ls} - x_{le}||x_{lw} - x_{le}|} \tag{4}$$

These joint angles did not change with the body orientation, and thus were view-invariant. In the 2AFC task, the model would pick the synthetic pose whose view-invariant representation was closer to that of the target natural pose. As a direct comparison, we also tested a 2D skeleton-based model, where the joint angles were calculated from projected 2D skeletons (see Figure 3 (c)) and were not viewpoint-invariant.

Results

Psychophysics

We found that pose matching accuracy generally decreased with increasing differences between target and comparison viewpoints (see Figure 4).

We further conducted sensory threshold analysis on the behavioral data from $Group_1$ and $Group_2$ separately. We grouped trial results by typicality of underlying 3D poses and informativeness of viewpoints in natural images, and we explored human performance differences for typical vs. atypical poses as well as informative vs. less informative views. For each condition and each level of viewpoint difference, we fitted psychometric functions for pose matching accuracy with respect to `noise_level`. Cumulative Gaussian was used to fit these functions with quickpsy package (Linares & López-Moliner, 2016). Nonparametric bootstrap was applied to get statistics on sensory thresholds.

Figure 5 (a) presents $Group_1$ results with plots of human sensory thresholds vs. differences in target and comparison viewpoints, along with linear regression results (see dashed lines) for each condition. We found that sensory thresholds increased with increasing viewpoint differences for all conditions. For the comparison between typical and atypical poses, we performed Welch's t-test with alternative hypothesis that the mean of sensory thresholds for atypical poses were greater than that for typical poses. Results (Figure 5 (a) top) showed that sensory thresholds for atypical poses were consistently and significantly greater than that for typical poses in most cases. For the comparison between informative and less informative viewpoints, we also performed Welch's t-test with alternative hypothesis that the true difference in means was not equal to 0. However, we found no significant and consistent differences this time (Figure 5 (a) bottom). To rule out the possibility that the performance differences were due to subjects spending different amount of time for different conditions, we looked into subjects' median reaction time and found no substantial differences across different conditions.

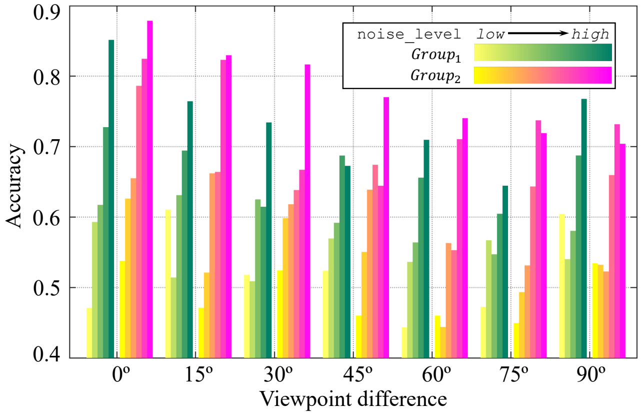


Figure 4: Pose matching accuracy (human) under different `noise_level` and viewpoint differences.

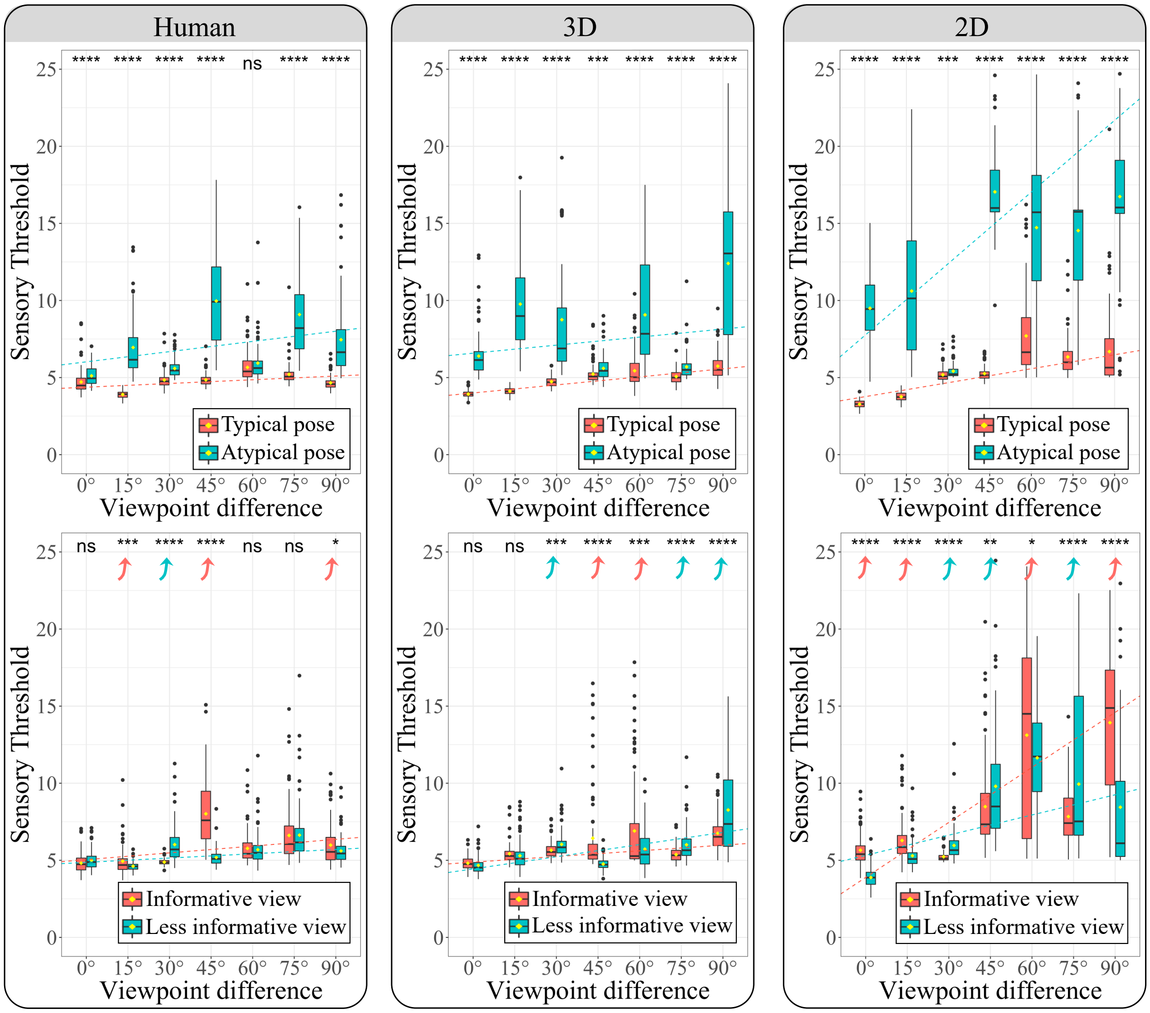
To validate these findings, we conducted the same analysis on $Group_2$ data, and obtained consistent results with $Group_1$.

Model Comparisons

We tested the aforementioned 2D and 3D skeleton-based pose matching models on all trials that human participants previously went through. For 2D and 3D model testing data, we performed the same sensory threshold analysis that was initially done on $Group_1$ and $Group_2$ behavioral data.

Figure 5 (b-c) shows model results on $Group_1$ trials. Sensory thresholds increased with increasing differences between target and comparison viewpoints for all conditions in both 2D and 3D skeleton-based models. In most cases, the slope of each regression line in 2D models was larger than its respective slopes in 3D models and humans. For the comparison of typical vs. atypical poses, sensory thresholds for atypical poses were significantly greater than that for typical poses in both models. We further compared the sensory threshold for skeleton-based models with the human sensory threshold. In typical pose condition (Figure 5 dashed red regression lines on the top), sensory thresholds for both models were comparable with that for humans. In atypical pose condition (Figure 5 dashed green regression lines on the top), however, sensory threshold for the 2D model was much larger than that for humans, while sensory threshold for the 3D model was more comparable with that for humans. For the comparison of informative vs. less informative viewpoints, we examined the results from both 2D and 3D skeleton-based models but found no consistent and significant differences.

Again to validate these findings, we conducted the same analysis using model testing data on $Group_2$ trials, and the



(a) (b) (c)

Figure 5: *Group*$_1$ sensory threshold results for humans, 3D and 2D pose matching models with linear regressions (dashed lines). Red and green up arrows in bottom plots indicate larger mean sensory thresholds for informative and less informative viewpoints respectively.

results were consistent with those from *Group*$_1$.

Discussion

Using the proposed 2AFC pose matching task, we were able to quantitatively measure the ability of humans to interpret 3D poses from natural images. We found that human performance on matching 3D poses decreased with increasing viewpoint differences between target and comparison poses. When comparing the slopes of the sensory threshold regression lines in the 2D model and humans, human performance decreased much slower, suggesting that they were unlikely to rely on pure 2D information for the task.

With these quantitative measurements, we first showed that this ability to interpret and match 3D poses depends on the typicality of the underlying 3D poses in natural images. Human performance for typical poses was significantly and consistently better than that for atypical poses, suggesting that humans may have more prior knowledge of typical pose structures. We also observed that both 2D and 3D models performed comparably to humans for typical poses. Nevertheless, for atypical poses, 3D model performance was much better than 2D model performance and was closer to human performance. This shows that 3D body knowledge is particularly important when interpreting images of atypical poses. The greater ability of human observers to match atypical poses, in contrast to the 2D model, is consistent with prior knowledge of 3D body structures. Thus, humans probably use such 3D body knowledge, at least for atypical pose matching.

Second, our results showed that human performance did not differ for different informativeness of viewpoints in natural images. As previous studies indicated, body representations in humans may have more viewpoint invariance (Sekunova, Black, Parkinson, & Barton, 2013). Hence one interpretation of our result is that humans may be robust to the change of viewpoints with different amount of body parsing information. However, it is worth noting that our objective measurements are primarily based on visibility of body joints from a viewpoint, which may not capture all aspects about the information supporting body parsing. The lack of performance difference could be due to human vision robustness to missing joints from minor or self occlusion in these natural images without major occluders. It is also possible that humans may recognize a joint with only a few pixels from less informative viewpoints, provided that local image regions contain minimal configurations for recognition. Future directions to address these issues include (1) exploring better ways to measure viewpoint informativeness and (2) adding systematically controlled occlusions onto target natural pose images in our task.

For model comparisons, the proposed pose matching task can be used to test the off-the-shelf pose estimation model on both typical and atypical poses from either informative or less informative viewpoints. Our psychophysical results provide useful benchmarks and comparisons to human performance on interpreting 3D poses from natural images across different pose and viewpoint conditions.

Conclusion

We designed a novel 2AFC pose matching task to quantitatively measure human ability to interpret 3D human poses in natural images. We showed that this human ability depends on pose typicality but not viewpoint informativeness defined by visibility of joints. By testing 2D and 3D pose matching models on the same task, we further showed that 3D body knowledge is particularly important when interpreting images of atypical poses, and that humans probably use prior knowledge of 3D pose structures.

Acknowledgments

We would like to thank Stephen Engel for helpful comments and suggestions on an early draft of this paper. This work was supported by National Institutes of Health with grant NIH R01 EY029700.

References

Abassi, E., & Papeo, L. (2020). The representation of two-body shapes in the human visual cortex. *Journal of Neuroscience*, *40*(4), 852–863.

Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3686–3693).

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, *94*(2), 115.

Dantone, M., Gall, J., Leistner, C., & Van Gool, L. (2014). Body parts dependent joint regressors for human pose estimation in still images. *IEEE Transactions on pattern analysis and machine intelligence*, *36*(11), 2131–2143.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, *293*(5539), 2470–2473.

Huynh, D. Q. (2009). Metrics for 3d rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision*, *35*(2), 155–164.

Johnson, S., & Everingham, M. (2011). Learning effective human pose estimation from inaccurate annotation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1465–1472).

Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M. J., & Gehler, P. V. (2017). Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6050–6059).

Lawson, R. P., Clifford, C. W., & Calder, A. J. (2009). About turn: The visual representation of human body orientation revealed by adaptation. *Psychological Science*, *20*(3), 363–371.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).

Linares, D., & López-Moliner, J. (2016). quickpsy: An r package to fit psychometric functions for multiple groups. *The R Journal, 2016, vol. 8, num. 1, p. 122-131.*

Liu, Z., Knill, D. C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision research*, *35*(4), 549–568.

Ma, Y., Paterson, H. M., & Pollick, F. E. (2006). A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior research methods*, *38*(1), 134–141.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *200*(1140), 269–294.

Mather, G., & Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *258*(1353), 273–279.

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., & Theobalt, C. (2018). Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 international conference on 3d vision (3dv)* (pp. 120–130).

Moon, G., Chang, J. Y., & Lee, K. M. (2019). Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 10133–10142).

Papeo, L., Stein, T., & Soto-Faraco, S. (2017). The two-body inversion effect. *Psychological science*, *28*(3), 369–379.

Peelen, M. V., & Downing, P. E. (2005). Selectivity for the human body in the fusiform gyrus. *Journal of neurophysiology*, *93*(1), 603–608.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, *343*(6255), 263.

Sekunova, A., Black, M., Parkinson, L., & Barton, J. J. (2013). Viewpoint and pose in body-form adaptation. *Perception*, *42*(2), 176–186.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive psychology*, *21*(2), 233–282.

Troje, N. F., & Westhoff, C. (2006). The inversion effect in biological motion perception: Evidence for a "life detector"? *Current biology*, *16*(8), 821–824.

Ullman, S. (1989). Aligning pictorial descriptions: An approach to object recognition. *Cognition*, *32*(3), 193–254.