Three-dimensional Pose Discrimination in Natural Images of Humans

Hongru Zhu[1] **Alan Yuille**[1] **Daniel Kersten**[2]
[1]Department of Cognitive Science, The Johns Hopkins University, Baltimore, MD 21218 USA
[2]Department of Psychology, University of Minnesota, Minneapolis, MN 55455 USA
hzhu38@jh.edu ayuille1@jh.edu kersten@umn.edu

Abstract

Recognizing and representing three-dimensional complex stimuli is an immense computational challenge our visual system is faced with. Much work has been done on the recognition of three-dimensional objects and the degree to which humans are able to interpret three-dimensional structures of complex objects. In this paper, we investigate this question with a novel method and prevalent set of stimuli — the human body in the natural world. Understanding the three-dimensional body pose structures is essential for more complex visual tasks like recognizing actions and social interactions. We investigated how well humans are able to interpret body poses in terms of their three-dimensional structures with a pose same-different matching task. We also examined whether this ability to perceive 3D poses depends on a priori knowledge of 3D body structure and the information provided in the projected image. Our results showed that humans' ability to match three-dimensional poses decreased with increasing viewpoint differences and that humans' performance depended on the typicality of the underlying three-dimensional pose. For typical poses which are presumably more familiar to subjects, performance is consistently better than atypical poses in most cases.

Keywords: pose estimation; same-different matching; three-dimensions

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 in that the two-dimensional projected images of the same 3D object vary considerably as 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. Much of the early focus in behavioral and neural studies focused on the problem of recognizing simple individual objects given viewpoint variation. Some theories have postulated that objects are represented with viewpoint-invariant part descriptions (Marr & Nishihara, 1978; Biederman, 1987) or view-specific three-dimensional representations in conjunction with normalization (Ullman, 1989). Other theories depended on multiple two-dimensional view-dependent representations, like view interpolations (Tarr & Pinker, 1989; Poggio & Edelman, 1990). There have been many studies as well as debates on the degree to which three-dimensional model-based information is used for object representation. Relatedly, recent computer vision researches have been attempting to address the debate by combining object recognition with three-dimensional object structures (Wu et al., 2017, 2018). However, there has been much less behavioral research on the problem of recognizing 3D objects from natural images, where the range of image variations is considerably larger. This study addresses the problem of recognizing human pose in natural images where in addition to the problem of viewpoint, vision needs to deal with self-occlusion, joint articulations as well as material changes due to clothing.

The human body is a stimulus that occurs quite frequently in daily life and carries a great deal of important information. Our visual system has developed dedicated mechanisms for processing body stimuli. Many studies have been done on human bodies to reveal perceptual and neural dimensions of human body representations. Several behavioral studies have demonstrated a high degree of sensitivity to properties like gender, mood, identity, etc (Ma, Paterson, & Pollick, 2006; Mather & Murdoch, 1994; Pollick, Lestou, Ryu, & Cho, 2002; Troje & Westhoff, 2006). Other researchers have investigated event related potential (ERP) components underlying human body detection (Taylor, Roberts, Downing, & Thierry, 2010). Finally, brain regions that are selective for human bodies have been identified and several theories as to their organization and function structures have been proposed (Downing, Jiang, Shuman, & Kanwisher, 2001; Peelen & Downing, 2005). While a wealth of research has been conducted on these subjects, we still have limited understanding when it comes to humans' ability to interpret body poses in three dimensions.

This study aims to investigate to what degree humans can interpret body poses in terms of their three-dimensional structures. We examined this by a pose same-different matching task between natural pose images and synthetic pose images. We tested whether subjects were able to match two human body images with the same underlying pose — one is from a natural image while the other is a synthetic body image generated with the same pose parameter but using different textures/clothing from a different viewpoint. Texture differences in natural and synthetic images allow us to measure the effect of appearance changes on humans' ability to perceive three-dimensional poses. Appearances of synthetic bodies depended on viewpoint and rendering, holding material and lighting constant. Thus, key comparisons focused on three-dimensional poses under different viewpoints. If hu-

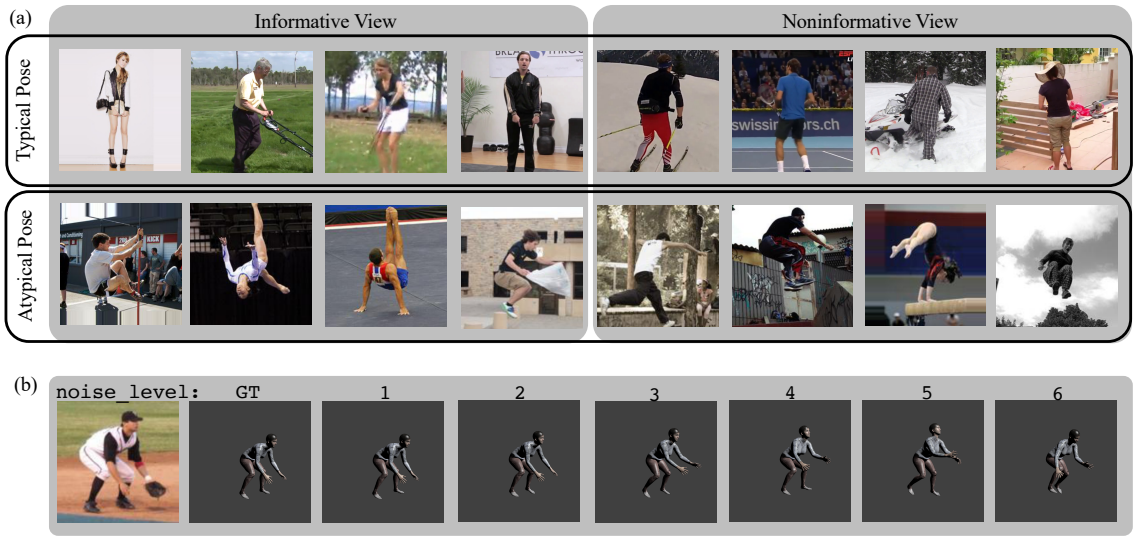

Figure 1: (a) Examples of selected images used in our pose matching task from different categories. (b) An example image (leftmost) with its noise-free synthetic pose (second left) as well as synthetic poses generated with different `noise_level`.

mans had no knowledge of three-dimensional pose structures, they would hardly be able to matching poses from different viewpoints. Otherwise, they should be able to perform the task to some extent. It is also unlikely that they can do it as good as under the same viewpoint. 3D bodies can take many different shapes and it is unlikely for people to have knowledge about all of them and apply the correct underlying shape for each image. Our prediction was that people can match poses but the performance would drop when the viewpoint differences between natural and synthetic images became larger.

We also examined the degree to which humans' ability to interpret three-dimensional structures of body poses depends on a priori knowledge of 3D body structure and the information provided in the projected image. From the literatures for the recognition of 3D objects, converging evidence has shown that humans' performance depends on familiarity of the object's appearance and views (Rock & DiVita, 1987; Tarr & Pinker, 1990; Bülthoff & Edelman, 1992; Liu & Kersten, 1998). In the context of human body pose, we selected two predictors — pose typicality and viewpoint informativeness. Despite various ways to measure these two factors, in this paper, we defined them as:

1. Pose typicality, which measures whether a 3D pose is typical (presumably more familiar to humans) or atypical (presumably less familiar to humans) in the dataset.
2. Viewpoint informativeness, which measures the degree to which the projected image is informative, i.e. with more visible joints or noninformative with more occluded joints.

We expected subjects' performance to be different under different predictor conditions, which would help us to understand what factors would influence humans' ability to perceive three-dimensional pose structures.

Method

Participants

Two groups of Amazon Mechanical Turk workers based in US participated ($n_1 = 35, n_2 = 42$). This study was approved by JHU IRB `HIRB00007053`.

Stimuli

Natural human body images 400 human pose stimuli were collected from *Unite the People* Dataset (Lassner et al., 2017). We built objective measurements regarding pose typicality and viewpoint informativeness respectively on this dataset (See Appendix A.1 for more details). Generally speaking, the objective measurement of pose typicality was based on how similar the 3D pose is to all the other poses in the dataset. The objective measurement of viewpoint informativeness was based on how many parts were visible from this viewpoint and how large those parts were in natural images. With these two measures, we quantitatively divided *Unite the People* Dataset into these four categories:

`Typical pose, Informative viewpoint`
`Typical pose, Noninformative viewpoint`
`Atypical pose, Informative viewpoint`
`Atypical pose, Noninformative viewpoint`

We randomly sampled 100 stimuli from each category. Then we split the 400 images into two groups of 200 images ($Group_1$ and $Group_2$) and tested them on two groups of subjects separately. Figure 1 (a) shows example of selected images from the four categories.

Synthetic human body images From *Unite the People* Dataset, we obtained 3D body joint rotation parameters for human bodies in each natural image, and used Blender 2.79 to make 2D projections of 3D posed synthetic humans.

Figure 2: An example experiment trial where the noisy synthetic poses (bottom right) are generated with `noise_level`= 5 and both noise-free (bottom left) and noisy synthetic poses was rotated with `rotation_level`= 2 (i.e. horizontally rotated by 30 degrees from the original viewpoint).

For the 2AFC pose matching task, we created both noise-free and noisy synthetic humans with same appearances but different poses. To create an synthetic human with joint rotation noises, we add $(2\times$`noise_lvl`$-1)\pi/128$ with a random sign $\{+,-\}$ to each rotation angle in axis-angle representation for each body joint. Same `noise_level` was applied to all joints in a pose. That is, we changed the relative 3D rotation of all joints with respect to their parents in the kinematic tree. Noise levels were determined by pilot experiments. For $Group_1$, `noise_level` was from 1 to 5, and for $Group_2$ we had an additional `noise_level` 6. Figure 1 (b) shows an example image with synthetic poses generated with different `noise_level`.

Once we get all pairs of synthetic humans with and without joint rotation noises at each `noise_level`, we also rotated the whole body for each pair horizontally by $15\times$`rotation_level` degrees so that viewpoints were changed from the original viewpoint in natural images. `rotation_level` was from 0 to 6.

Procedure

Participants were divided into two groups. Each subject went through 200 trials with natural images from either $Group_1$ or $Group_2$ shown only once across the experiment.

Each subject went through three practice trials before they started the experiment. In each trial, they were shown a screen with three images. On the top was one natural image of a human. On the bottom are two synthetic humans. One was rendered without any additive joint rotation noises while the other was rendered with noises at a random `noise_level`. Further, the two synthetic human images were rotated to a random `rotation_level` before rendered into two-dimensional projections. In each trial, these two synthetic human images were randomly placed to the left/right of the screen. Figure 2 shows an example trial when `noise_level`= 5 and `rotation_level`= 2. Subjects were asked to judge the human pose in the natural and synthetic images, and respond by selecting the synthetic human whose pose matches the pose in the natural image better, regardless of viewpoint differences introduced by whole body rotations. Subjects were given unlimited time to respond to each trial while the three images were constantly shown on the screen. Subject can choose to take a break after finishing 100 trials.

Data preprocessing

After obtained experiment data from all subjects, we went through two preprocessing stages to make sure that our analysis was conducted based on reliable data.

First, we excluded trials where either (1) the noise-free synthetic pose was not a good fit to the body pose in the corresponding image (with very large fitting errors) or (2) the noisy synthetic poses were illegal (e.g. arms going through bodies).

Second, we disregarded data from unattentive subjects with a threshold accuracy for those trials where noisy synthetic poses at the highest `noise_level` was used with no viewpoint differences i.e. `rotation_level`= 0. These ought to be the easiest trials and we expected subjects to perform well on them. The threshold accuracy was set to be 0.55 for `noise_level`= 5 in $Group_1$ and 0.70 for `noise_level`= 6 in $Group_2$.

After these two preprocessing stages, we had 5,040 trials from 28 subjects in $Group_1$ and 6,130 trials from 33 subjects in $Group_2$. We used these trials for further data analysis.

Results

Accuracy vs noise and viewpoint changes

To see how accuracy changes when the noisy synthetic humans were at different `noise_level`, we plotted pose matching accuracy vs `noise_level` for different levels of viewpoint changes in Figure 3.

On the one hand, accuracy went up as `noise_level` increased, which makes sense considering that the noise-free synthetic poses would be more dissimilar to its noisy counterparts. On the other hand, accuracy tended to drop, but not sharply to the chance level, when viewpoint differences increased from 0 to 90 degrees. It remains unknown that (1) to what extent humans can perceive three dimensional pose structures with viewpoint changes and (2) whether this ability depends on knowledge of 3D body structure and the information provided in the projected image. To answer these questions, we compare accuracy results using different predictors and showed further results below.

Typical vs atypical pose

We divided trials into two sets based on the pose typicality in natural images and expected this can be a predictor for human's ability to perceive three-dimensional pose structures. Specifically, we would like to see to what degree

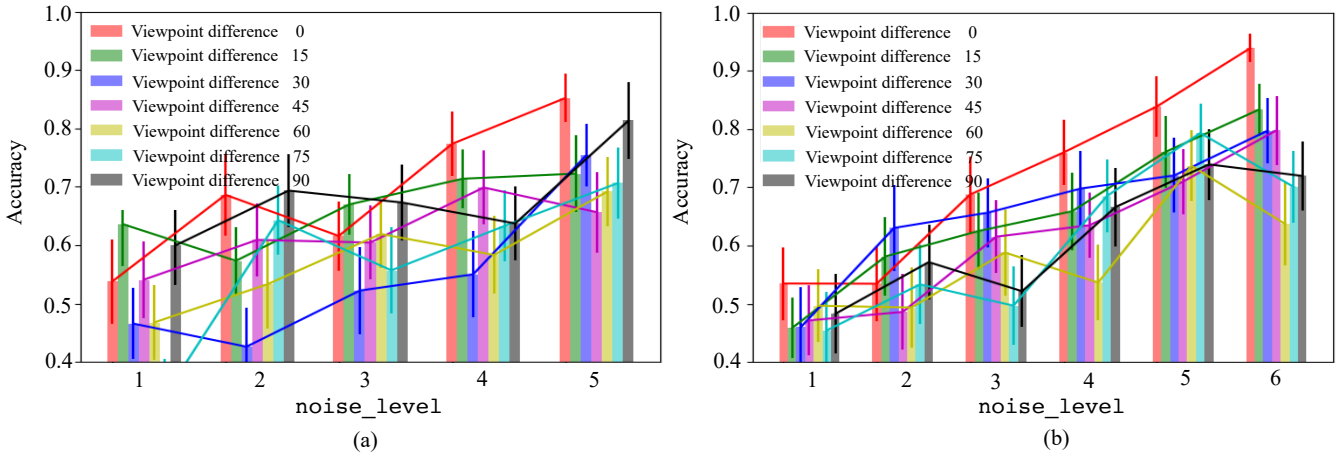

Figure 3: Accuracy for pose matching with viewpoint changes when using noisy synthetic poses from different `noise_level`. Results from two independent image groups $Group_1$ (a) and $Group_2$ (b) are shown separately.

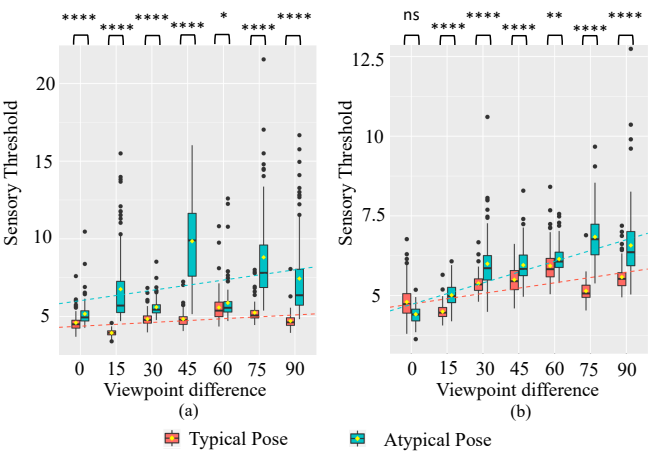

Figure 4: Sensory thresholds for typical and atypical poses. Results from two independent image groups $Group_1$ (a) and $Group_2$ (b) are shown separately.

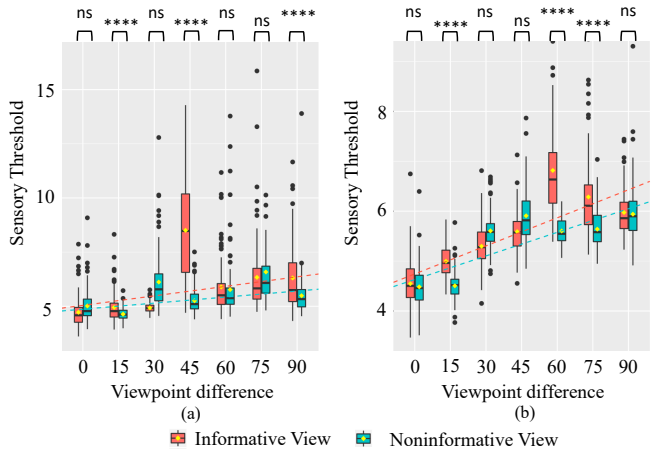

Figure 5: Sensory thresholds for informative and noninformative viewpoints. Results from two independent image groups $Group_1$ (a) and $Group_2$ (b) are shown separately.

humans were able to match poses with viewpoint changes and whether the degree was different for typical and atypical poses. Therefore, we fitted psychometric functions for pose matching accuracy w.r.t. `noise_level` under different levels of viewpoint changes. A cumulative Gaussian was used to fit these functions with quickpsy package (Linares & López-Moliner, 2016). A nonparametric bootstrap was applied to get statistics on the sensory thresholds.

Figure 4 plots sensory thresholds vs different levels of viewpoint changes between target natural images and synthetic ones. For both typical and atypical pose categories, sensory thresholds got larger with increased viewpoint differences, indicating that humans' ability to match poses were more compromised when viewpoint differences were large. In addition, Welch's t-test was performed to determine whether sensory thresholds for atypical poses were significantly GREATER than typical poses at each level of viewpoint changes. As Figure 4 shows, sensory thresholds for atypical pose category were significantly greater than typical pose category in most cases. Results were consistent for the two independent image groups.

To rule out the uninteresting hypothesis that these differences for typical and atypical poses were due to subjects spending different amount of efforts on different images, we plotted the median reaction time for atypical and typical pose categories in Figure 6 (see Appendix A.2), which showed that subjects' reaction time was not substantially different. We also tried testing a different predictor — viewpoint informativeness, the results of which are shown below.

Informative vs noninformative viewpoint

We divided our trials into informative and noninformative viewpoint categories, and conducted similar analysis as done above for typical vs atypical poses. We fitted psychometric functions for accuracy w.r.t. `noise_level` under different levels of viewpoint changes and plotted the sensory thresholds in Figure 5. Again, sensory thresholds were increasing as viewpoint differences increased for both categories. Welch's t-test were performed to determine whether sensory thresholds were significantly DIFFERENT for informative and noninformative viewpoint categories. However, we did not see significant differences this time. Nor does the median reaction time shown in Figure 6 in Appendix A.2.

Discussion

From two independent image sets, we had consistent results about humans' ability to match body poses from different viewpoints. With the increased viewpoint differences between natural and synthetic images, the sensory thresholds slowly grew larger. There are at least two possible explanations for this slowly decreased performance level. One is that people may have some knowledge of three-dimensional body structures so that their performance did not drop sharply with viewpoint changes. It is also possible that subjects picked up purely 2D information from poses with different appearances and viewpoints, and their template matching performance decreased due to increased image differences. To distinguish these two alternatives, more analysis would be required as future work such as to design 3D ideal observers and fit human performance.

Results from contrasting typical poses with atypical poses showed a significant difference in terms of accuracy. Subjects performed better for matching typical poses in most cases, suggesting that they may potentially have better knowledge of the three-dimensional structure for typical poses than for atypical poses. Given that typical poses are presumably more familiar to humans, this finding is similar to those findings for object recognition where humans' performance depended on familiarity of the object appearances.

When looking into the comparison from the viewpoint informativeness predictor, there was no significant differences between the two viewpoint categories. Humans might be surprisingly robust to different viewpoint informativeness. However, it could also be that humans are not actually that robust and that viewpoint is still a predictor for humans' ability to interpret body poses in three dimensions. On the one hand, it could be that our measure of viewpoint informativeness (see appendix A.1) may not capture the essential features as familiar viewpoints do, and there is not a large margin between the viewpoint informative score distribution for informative vs. noninformative viewpoint images (Figure 7). One future direction is to better characterize how much information people can get from a viewpoint, possibly with an ideal observer model. On the other hand, it also could be due to the way we test pose matching across different viewpoints. Even though we measured and controlled viewpoint informativeness in target natural images, we did not take into account nor analyze the viewpoint informativeness of synthetic images. Therefore, it could happen that in some trials, natural images were from an informative viewpoint while synthetic images were from a noninformative viewpoint. Subjects might be able to perceive three dimensional structures for natural poses but not for synthetic poses, and thus be unable to do pose matching. Therefore, another future direction is to take into account viewpoint informativeness in both natural and synthetic images and evaluate the effectiveness of viewpoint on predicting humans' performance on pose matching.

There are of course other directions worth exploring. This work provides a paradigm to test humans' ability to interpret poses in three dimensions. A natural followup is to test pose matching with appearance changes introduced by occluders. Occlusion could be another predictor and potentially reveal the robustness of humans' ability to interpret poses in three-dimensions. Another direction is to investigate computational models that are able to perform such pose matching tasks. Like model that can do face identification across different viewpoints, models that learn to do pose matching across different viewpoints in natural images may provide more insights on what features are essential for interpreting poses in term of their three-dimensional structures.

In conclusion, we found consistent results suggesting that humans are able to interpret body poses in terms of their three-dimensional structures but not without any errors or ambiguities. Similar to the recognition of three-dimensional objects, humans' performance is better for typical poses, which are presumably more familiar to humans. Future work is needed to better reveal and understand the effect of viewpoint informativeness and other possible predictors on humans' ability to interpret body poses in three dimensions.

References

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.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, *89*(1), 60–64.

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. (2010). Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc* (Vol. 2, p. 5).

Johnson, S., & Everingham, M. (2011). Learning effective human pose estimation from inaccurate annotation. In *Cvpr 2011* (pp. 1465–1472).

Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M. J., & Gehler, P. V. (2017, July). Unite the people: Closing the loop between 3d and 2d human representations. In *Ieee conf. on computer vision and pattern recognition (cvpr)*. Retrieved from `http://up.is.tuebingen.mpg.de`

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., & Kersten, D. (1998). 2d observers for human 3d object recognition? In *Advances in neural information processing systems* (pp. 829–835).

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015, October). SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, *34*(6), 248:1–248:16.

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.

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.

Pollick, F. E., Lestou, V., Ryu, J., & Cho, S.-B. (2002). Estimating the efficiency of recognizing gender and affect from biological motion. *Vision research*, *42*(20), 2345–2355.

Rock, I., & DiVita, J. (1987). A case of viewer-centered object perception. *Cognitive psychology*, *19*(2), 280–293.

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

Tarr, M. J., & Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science*, *1*(4), 253–256.

Taylor, J. C., Roberts, M. V., Downing, P. E., & Thierry, G. (2010). Functional characterisation of the extrastriate body area based on the n1 erp component. *Brain and cognition*, *73*(3), 153–159.

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.

Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, B., & Tenenbaum, J. (2017). Marrnet: 3d shape reconstruction via 2.5 d sketches. In *Advances in neural information processing systems* (pp. 540–550).

Wu, J., Xue, T., Lim, J. J., Tian, Y., Tenenbaum, J. B., Torralba, A., & Freeman, W. T. (2018). 3d interpreter networks for viewer-centered wireframe modeling. *International Journal of Computer Vision*, *126*(9), 1009–1026.

Appendix

A.1 Pose and viewpoint measurements

We built objective measurements of pose typicality and viewpoint informativeness for *Unite the People* Dataset to help us select 400 human body images used in our behavioral study. We assigned a pose typicality score and a viewpoint informative score to each image.

Pose typicality score *Unite the People* Dataset consists of 8,515 images from four datasets: *LSP (Johnson & Everingham, 2010), LSP-extended (Johnson & Everingham, 2011), MPII (Andriluka, Pishchulin, Gehler, & Schiele, 2014)* and *FashionPose (Dantone, Gall, Leistner, & Van Gool, 2014)*. To find more reliable atypical pose images from the whole dataset, each image was compared with all other pose images from *Unite the People* Dataset.

When comparing two pose images, we calculated pose distances by comparing their body joint rotation angles. *Unite the People* Dataset provides 3d rotation angles for all body joints in axis-angle format. Following (Loper, Mahmood, Romero, Pons-Moll, & Black, 2015), these axis-angles represent the relative rotation of one joint with respect to its parent in the kinematic tree. We first convert axis-angle representations into unit quaternions. Body pose i with m joints is represented by a list of quaternions $[q_{i1}, q_{i2}, ...q_{im}]$. Then distance between body i and body j is:

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

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). Thus typicality score for a given body pose i can be calculated by:

$$\frac{\sum_{j \in U} D(i,j)}{size(U)} \tag{2}$$

Where U is the set of images from *Unite the People* Dataset excluding image i.

Figure 7 (left) shows the pose typicality score distribution for our selected 400 images divided into typical pose category

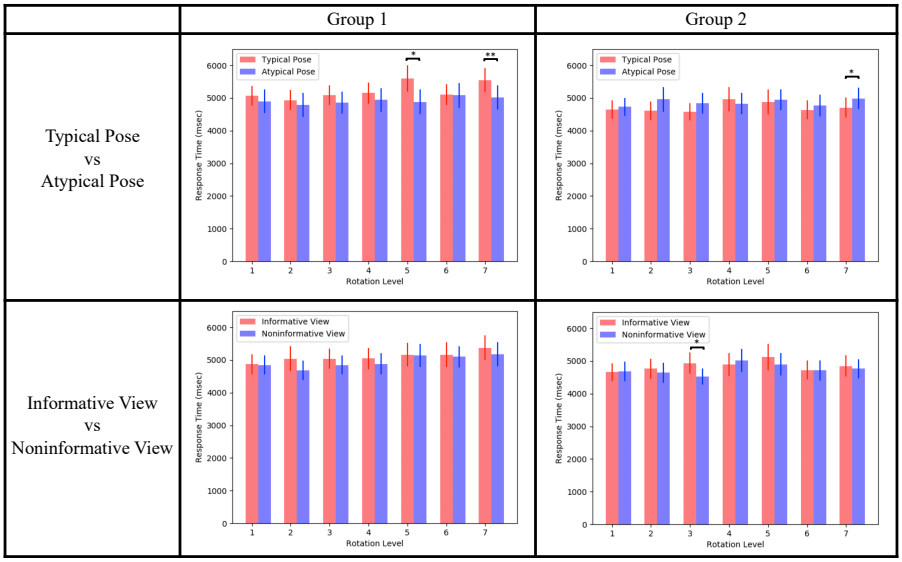

Figure 6: Median reaction time (RT) for images from $Group_1$ and $Group_2$. The first row compared median RT for typical and atypical pose image trials across different `rotation_level`. The second row compared median RT for informative and noninformative viewpoint image trials.

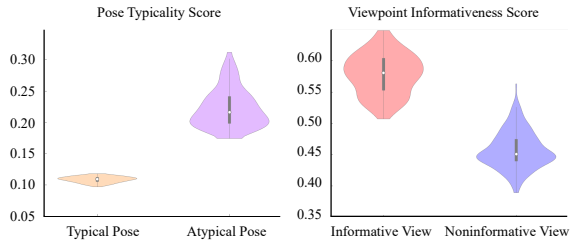

Figure 7: Left: Pose typicality score distribution for typical and atypical pose images. Right: Viewpoint informativeness score distribution for informative and noninformative viewpoint images.

and atypical pose category. A higher score means that the pose is more distant to other images from the dataset, making it more likely to be an atypical pose.

Viewpoint informativeness score To quantitatively measure how informative the viewpoint is for a given image, we used the segmentation mask provided by *Unite the People* Dataset and calculated the fraction of joint pixels for 14 joints defined in *LSP* w.r.t the whole body. We averaged these fractions and use this as an indicator for informativeness. Given that images from *Unite the People* Dataset are cropped so that human bodies are reasonably large and roughly in the center, we did include the fraction of body pixels w.r.t all image pixels when calculating viewpoint informativeness score.

Specifically, for each body pose i, we calculate the fraction of pixels for joint k w.r.t the whole body as

$$f_{ik} = \frac{\#\text{ of pixels for joint } k \text{ of body } i}{\#\text{ of pixels for whole body } i} \tag{3}$$

Before averaging across different joints to get viewpoint informativeness score for body pose i, we standardized the fractions for different joints across all images in the dataset so that the fraction for each joint is from a standard normal distribution. This was done by calculating z-score:

$$\text{z-score}(f_{ik}) = \frac{f_{ik} - \mu_k}{\sigma_k} \tag{4}$$

where μ_k and σ_k are mean and standard deviation for f_{ik} for all body i from the dataset.

Thus the viewpoint informativeness score for body pose i is calculated as:

$$\frac{\sum_{k=1}^{14} \text{sigmoid}(\text{z-score}(f_{ik}))}{14} \tag{5}$$

Sigmoid is used to make sure scores fall between 0 and 1.

Figure 7 (right) shows the viewpoint informativeness score distribution for our selected 400 images divided into informative viewpoint and noninformative viewpoint category. A higher score means that more joints are visible with larger areas and therefore the pose is likely to be more informative than a pose whose joints are barely visible or occluded.

A.2 Pose matching reaction time

Subjects were given unlimited time to response to each trial and sometimes they may spend extremely long time on a trial (possibly taking a break). Therefore, we used the median reaction time instead of mean reaction time under each `rotation_level` condition for each subject. We plotted the averaged median reaction time over all subjects in Figure 6.