Pictorial Human Spaces: How Well do Humans Perceive a 3D Articulated Pose?

*Elisabeta Marinoiu[1], Dragos Papava[1], Cristian Sminchisescu[2,1]
[1]Institute of Mathematics of the Romanian Academy
[2]Department of Mathematics, Faculty of Engineering, Lund University
elisabeta.marinoiu@imar.ro, dragos.papava@imar.ro, cristian.sminchisescu@math.lth.se

Abstract

Human motion analysis in images and video is a central computer vision problem. Yet, there are no studies that reveal how humans perceive other people in images and how accurate they are. In this paper we aim to unveil some of the processing–as well as the levels of accuracy–involved in the 3D perception of people from images by assessing the human performance. Our contributions are: (1) the construction of an experimental apparatus that relates perception and measurement, in particular the visual and kinematic performance with respect to 3D ground truth when the human subject is presented an image of a person in a given pose; (2) the creation of a dataset containing images, articulated 2D and 3D pose ground truth, as well as synchronized eye movement recordings of human subjects, shown a variety of human body configurations, both easy and difficult, as well as their 're-enacted' 3D poses; (3) quantitative analysis revealing the human performance in 3D pose reenactment tasks, the degree of stability in the visual fixation patterns of human subjects, and the way it correlates with different poses. We also discuss the implications of our findings for the construction of visual human sensing systems.

1. Introduction

When shown a photograph of a person, humans have a vivid, immediate sense of 3D pose awareness, and a rapid understanding of the subtle body language, personal attributes, or intentionality of that person. How can this happen and what do humans perceive? How is such paradoxical monocular stereoscopy possible? Are the resulting percepts accurate in an objective, veridical sense, or are they an inaccurate, possibly stable by-product of extensive prior interaction with the world, modulated by sensations acquired through the selective visual observation of the photograph? The distinction between the *regular* 3D space we move in and the 3D space perceived when looking *into* a photograph–the *pictorial space*–has been introduced and beautifully studied by Koenderink[11] for rigid objects through the notion of *pictorial relief*. In this paper we aim to explore the concept for the case of articulated human structures, motivating our *pictorial human space* terminology.[1] Our approach to establish the observation-perception link is to make humans re-enact the 3D pose of another person (for which ground truth is available), shown in a photograph, following a short exposure time of a few seconds. Simultaneously our apparatus allows the measurement of human pose and eye movements during the 'pose matching' performance. This comprises an observation, a memory and a re-enactment error. As the poses are taken from everyday activities, their reproduction should not put subjects in difficulty as far as the ability to articulate a pose is concerned, however. The contribution of our work can be summarized as follows: (1) the construction of an apparatus relating the human visual perception (re-enactment as well eye movement recordings) with 3D ground truth; (2) the creation of a dataset collected from 10 subjects (5 female and 5 male), containing 120 images of humans in different poses, both easy and difficult, available online at `http://vision.imar.ro/percept3d`; and (3) quantitative analysis of human eye movements, 3D pose reenactment performance, error levels, stability, correlation as well as cross-stimulus control, in order to reveal how different 3D configurations relate to the subject focus on certain features in images, in the context of the given task. We conclude with a discussion of the implications of our findings for the construction of 3D human pose analysis systems.

Related Work: The problem of human pose estimation from static images has received significant attention in computer vision, both in the 2D [7, 15, 26] and the 3D case[5, 17, 21, 24, 1, 8, 3, 2]. The 2D case is potentially easier, but occlusion and foreshortening challenge the generalization

*Authors contributed equally.

[1]No connection with pictorial structures[15]–2D tree-structured models for object detection.

ability of 2D models. For 3D inference, different models as well as image features have been explored, including joint positions[12, 4], edges and silhouettes[21, 8]. Recent studies focused, as well, on inferring human attributes extracted from 3D pose[18] and on analyzing perceptual invariance based on 3D body shape representations[16]. It is well understood that the problem of geometrically inferring a skeleton from *monocular* joint positions, the problem of fitting a volumetric model to image features by non-linear optimization, and the problem of predicting poses from a large training data corpus based on image descriptors are underconstrained under our present *savoire faire*. These produce either discrete sets of forward-backward ambiguities for known limb lengths[12, 21, 19] (continuous non-rigid affine folding for unknown lengths), or multiple solutions due to incorrect alignment or out-of-sample effects[5, 21]. 3D human pose ambiguities from monocular images may not be unavoidable. Better models and features, a subtle understanding of shadows, body proportions, clothing or differential foreshortening effects may all reduce uncertainty. The question still is whether such constraints can be reliably integrated towards metrically accurate monocular results. Here we take an experimental perspective, aiming to better understand what humans are able to do, how accurately, and where they are looking when recognizing a 3D pose. Such insights can have implications in defining more realistic targets and levels of uncertainty for the operation of computer systems in similar tasks, and could provide quantitative hints (and training data) on what image features to focus on to achieve such performance. While this work focuses on experimental human sensing in monocular images, the moving light display setup of Johansson[10] is worth mentioning as a milestone in emphasizing the sufficiency of dynamic minimalism with respect to human motion perception. Yet in that case, as for static images, the open question remains on how such vivid dynamic percepts relate to the veridical motion and how stable across observers they are. Our paper focuses mostly on analysis from a computer vision perspective but links with the broader domain of sensorimotor learning for redundant systems, under non-linearity, delays, uncertainty and noise[25]. We are not aware, however, of a study similar to ours, nor of an apparatus connecting real images of people, eye movement recordings and 3D perceptual pose data, with multiple subject ground truth.

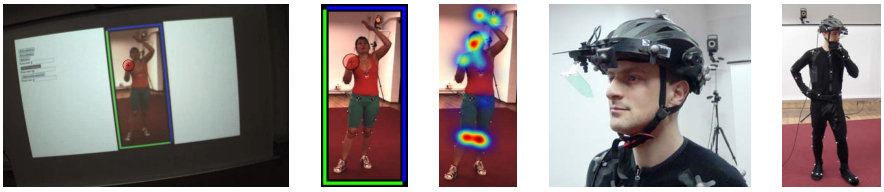
2. Apparatus for Human Pose Perception

The key difficulty in our experimental design is to link a partially subjective phenomenon like the 3D human visual perception with measurement. Our approach was to dress people in a motion capture suit, equip them with an eye tracker and show them images of other people in different poses, which were obtained using motion capture as well (fig. 1). By asking the subjects to re-enact the poses shown, we can link perception and measurement. We use a state-of-the-art Vicon motion capture system together with a head mounted, high-resolution mobile eye tracking system. The mocap system tracks a large number of markers attached to the full body mocap suit worn by a person. Each marker track is then labeled taking into account its placement on the body regarding a model template. Having all these labels, human models are used to accurately compute the location and orientation of each 3D body joint. The mobile eye tracker system maps a person's gaze trajectory on the video captured from its frontal camera. The synchronization between the two cameras is done automatically by the system.

2.1. Setup and Dataset Collection

We have analyzed the re-enactment performance of 10 subjects, 5 male and 5 female, who did not have a medical history of eye problems or mobility impediments. The participants were recruited through an agency and had no link with computer vision. Each subject was asked to look at images of people in different poses taken from the Human3.6M (H3.6M) dataset[9], and then reproduce the poses seen as well as they could. They were explicitly instructed not to mirror the pose, but to reproduce the left and right sides accordingly. The images were projected on a screen located 2.5–3 meters away from them and 1.2 meters tall. The eye tracker calibration was done by asking the subject to look at specific points, while the system was recording pupil positions at each point. The calibration points were projected on the same screen used to project images.

Each subject was required to stand still and look at one image at a time until it disappeared, then re-enact the pose by taking as much time as necessary. We chose to display images for 5 seconds such that the subjects would have enough time to see all the necessary pose detail, while still being short enough not to run into free viewing. The duration was chosen by first recording two test subjects and showing them images for 3, 5 and 8s. Their feedback was that 3s was too short to view enough detail, while 8s was more than enough. From an eye tracking video we were interested in the 5s that captured the projected pose along with the gaze recordings over the image of the person in that pose. To increase the accuracy of recorded gazes, we mapped fixations that fell onto the image on the screen (captured by eye-tracker/viewer's camera) back to the original high-resolution image (*c.f.* fig. 1). We created green and blue borders for the original image, to easily detect and track later on. First, we evaluated the viewer's camera intrinsic parameters, and corrected for radial distortion of each image in the captured video. Then we thresholded in the HSV color space to retrieve the green and blue borders. Instead of directly detecting corners (which due to subject's subtle head movement might fall-off the image),



(a) (b) (c) (d) (e)

Figure 1. Illustration of our human pose perception apparatus. (a) Screen on which the image is projected as captured by the external camera of the eye tracker. (b) Result of mapping the fixation distribution on the original high-resolution image, following border detection, tracking and alignment. (c) heat map distribution of all fixations of one of our subjects for this particular pose. (d) detail of our head-mounted eye tracker and (e) 3D motion capture setup.

we detected the green and blue borders using a Radon transform, imposing a 90° angle and a known aspect ratio of the image on the screen. Moreover, we synchronized the mocap and eye tracker system by detecting the start and end frame of the displayed images in the eye tracker video as well as in the video recorded with the motion capture system (two digital cameras of the mocap system were pointed at the screen as well). During experiments, subjects were dressed in the mocap suit and had the mobile eye tracker head-mounted. For each pose projected on the screen, we captured both the eye-tracks and the 3D movement of the subject in the process of re-enacting the pose, once it had disappeared from the screen.

Once the 5s exposure time passed, the subject no longer had the possibility to see the image of pose to re-enact, but had to adjust his position based on the memory of that pose. This time constraint ensures that the subject will mostly look at what is important in understanding and reproducing the pose. Moreover, it makes the process of translating fixations from the video coordinates of the eye tracker to the ones of the image on the screen robust, as there are no rapid head movements or frames where the pose is only partially seen (fig. 1). Another approach would have been to allow the subject to look at the image while adjusting his pose, thus removing a confounding factor due to short-term memory decay (forgetting). In this way, the subjects could alternate between adjusting body parts and checking back in the image, without the constraint to memorize all the pose details. While this choice may appear more natural, it has the drawback of no longer being able to easily map the human fixations to the presented image, and makes it difficult to separate fixations that fall on the pose from those on one's own body or the surroundings. This experiment, presented in our accompanying technical report[13] indicates that, perhaps surprisingly, this supplementary visual aid does not improve the pose re-enactment performance.

We display a total of 120 images, each representing a bounding box of a person, and rescaled them to 800px height in order to have the same projected size. The images are mainly frontal; 100 contain easily reproducible standing poses, whereas 20 of them are harder to re-enact as they require sitting on the floor, with often additional self-occlusion. The poses shown were selected from Human3.6M[9], among various types of daily activities and were performed by 10 different people.

2.2. Evaluation and Error Measures

We use the same skeleton joints as in Human3.6M[9] such that our analysis can immediately relate with existing computer vision methods and metrics.

H3.6M position error (MPJPE) between a recorded pose and the ground truth is computed by translating the root (pelvis) joint of the given pose to the one of the ground truth. We rotate the pose such that it faces the same direction as the ground truth. The error is then computed for each joint as the norm of the difference between the recorded pose and ground truth. In this way we compensate for the global orientation of the subject. We normalize both the subject skeleton and the ground truth to a default skeleton of average weight and height, ensuring that all computed errors are comparable between poses and subjects.

H3.6M angles error (MPJAE) is computed as the absolute value of the difference between joint angles of test and ground truth, for each 1d.o.f. joint (*e.g.*, for the elbow, the angle between upper arm and lower arm). For a 3d.o.f. joint, the representation is in ZXY Euler angles and the angle difference is computed separately for each d.o.f. as previously; the final error is the mean over the 3d.o.f. differences.

3. Data Analysis

3.1. Human Eye Movement Recordings

Static and Dynamic Consistency. In this section we analyze how consistent the subjects are in terms of their fixated image locations. We are first concerned with evaluating static consistency which considers only the location of the fixations and then dynamic consistency which takes into account the order of fixations. To evaluate how well the subjects agree on fixated image locations, we predict each subject's fixations in turn using the information from the other 9 subjects[6, 14]. This was done considering the same pose as well as different poses. For each pose, we generate

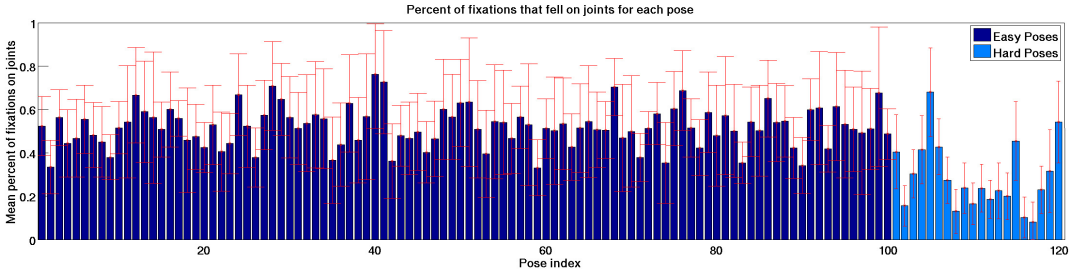


Figure 2. Percentage of fixations falling on joints, for each of the 120 poses (easy and hard) shown to subjects. The mean and standard deviation is computed for each pose among the 10 subjects. A fixation was considered to fall on a particular joint if this was within 40px distance from the fixation. On average 54% of the fixations on easy poses and 30% of those for hard poses fell on joints.

a probability distribution by assigning a value of 1 to each pixel fixated by at least one of the 9 subjects and 0 to others, then locally blurring with a Gaussian kernel. The width was chosen such that, for each pose, it would span a visual angle of 1.5^{o}. The probability at pixels where the 10^{th} subject's fixations fall is taken as the prediction of the model obtained from 9 subjects. For cross-stimulus control we repeat the process for 100 pairs of randomly selected different poses. Fig. 3 indicates good consistency.

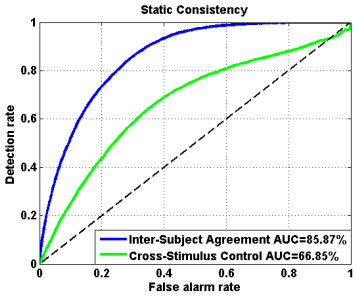


Figure 3. Static inter-subject eye movement agreement. Fixations from one subject are predicted using data from the other 9 subjects both on the same image (blue) and on a different image of a person, randomly selected from our 120 poses (green).

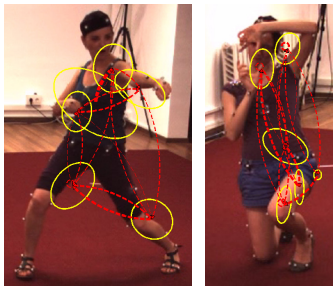


Figure 4. HMMs trained for two poses using the method of [14]. Ellipses correspond to states, whereas dotted arrows to transition probabilities assigned by the HMM. The AOIs determined by the model correspond to regions that well characterize the pose.

To evaluate how consistent the subjects are in their order of fixating areas of interest (AOIs), we used the hidden Markov modeling recently developed by [14]. The states correspond to AOIs that were fixated by subjects and the transitions correspond to saccades. For each pose, we learn a dynamic model from the scanpaths of 9 subjects and compute the likelihood of the 10^{th} subject's scanpath under the trained model. The leave-one-out process is repeated in turn for each subject and the likelihoods averaged. The average likelihood (normalized by the scanpath length) obtained is -9.38. Results are compared against the likelihood of randomly generated scanpaths. Specifically, for each pose we generate a random scanpath with the exception of the first fixation which was taken as the center of the image to account for central bias. Each random scanpath is evaluated against the model trained with subject fixations on that pose. The average likelihood of randomly generated trajectories is much smaller than the one of subject data, only -42.03, indicating that subjects are consistent in the order they fixate AOIs. Examples of trained HMMs are shown in fig. 4.

What percentage of fixations fall on joints? In order to understand where our subjects look and what are the most important body cues in re-enacting each 3D pose, we project the skeletal joint positions onto the image shown. We analyze the fixations relative to the 17 joints in fig. 5. For each pose, we take into account the 3D occlusions (based on mocap data and a 3D volumetric model) and consider, as possibly fixated, only those joints that are visible. We consider a fixation to be on a joint if it falls within a distance of 40 pixels. This threshold was chosen to account for an angle of approximately 1.5^{o} of visual acuity. Our first analysis aims to reveal to what extent subjects are fixating joints and how their particular choice of regions varies with the poses shown. Fig. 2 shows what percent of fixations fell on joints for each pose, for each subject. On average 54% of fixations fell on various body joints for easy poses, but only 30% for hard poses. This is not surprising as more joints are usually occluded in the case of the complex poses shown.

Where do subjects look first? Since approximately half of human fixations fell on joints, we want to know whether certain joints are always sought first, and to what extent the joints considered first are pose-dependent. The order

in which joints are fixated can offer insight into the cognitive process involved in pose recognition. Fig. 5 shows how many times a joint was among the first 3 AOIs fixated, for each subject. The first 3 fixations almost never fall on the lower body part which (typically) has less mobility, but mostly on the regions of the head and arms.

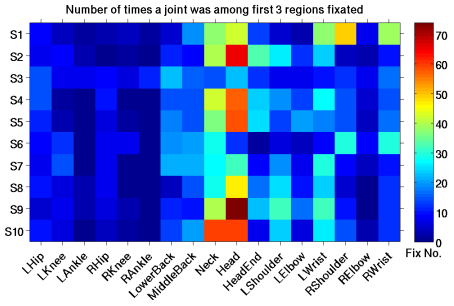


Figure 5. Number of times a body joint was in the top-3 regions fixated, accumulated over the 120 poses shown.

Which are the most fixated joints? We study whether certain joints are fixated more than others and we want to know whether this would happen regardless of the pose shown, or whether it varies with the pose. For this purpose, we consider the number of fixations that fall on a particular joint as well as the time spent on fixating each joint. Fig. 6 shows the distribution of fixations on body joints, averaged over poses. Notice that although certain joints have been fixated more than others, this depends on the specific pose. The variation can also be observed in a detailed analysis of how frequently certain joints were fixated in two arbitrary poses, presented in fig. 8 and fig. 9. While in fig. 8 the legs are almost never fixated, a different frequency pattern is apparent in fig. 9. One explanation could be that the leg positions in the second image have greater deviation form a standard, rest pose than in the first image. In fig. 7 we aggregate joint fixations for each subject, on all poses, and show the most frequently fixated joints, on average. The inter-subject variation is smaller than the one between poses, confirming a degree of subject consistency with respect to the joints more frequently fixated. The wrists and the head area are most looked at, within a general trend of fixating upper body parts more than lower ones.

How long are people looking at different joints? As the length of a fixation varies, it is also important to consider the time spent in fixating a particular joint. In fig. 10 we show the mean time and standard deviation spent on a pose, shown on each body joint, by aggregating over subjects. Similarly, in fig. 11 we show the mean time and standard deviation between subjects, by aggregating over poses. Notice that inter-pose standard deviation is higher than inter-subject standard deviation. It can be further observed that joints at the extremity of the body (head, neck, wrists) are the most fixated.

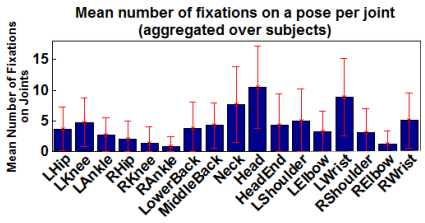


Figure 6. Fixation counts on each joint. The mean and standard deviation is computed among the 120 poses by aggregating over all 10 subjects, for each pose.

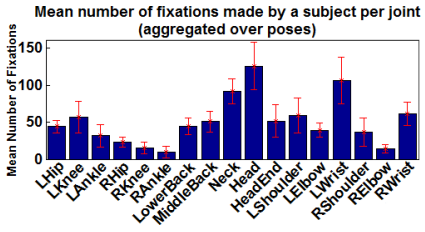


Figure 7. Fixation counts on each joint. The mean and standard deviation is computed for each of the 10 subjects by aggregating their fixations over all 120 poses.

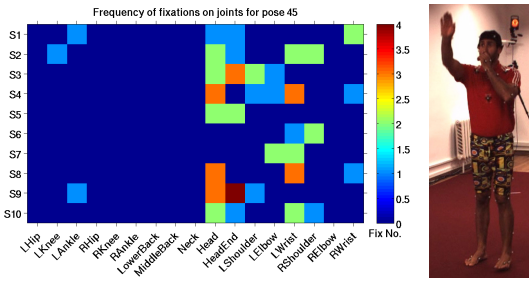


Figure 8. Number of fixations of each subject on the 17 body joints, when presented the pose-image shown on the right.

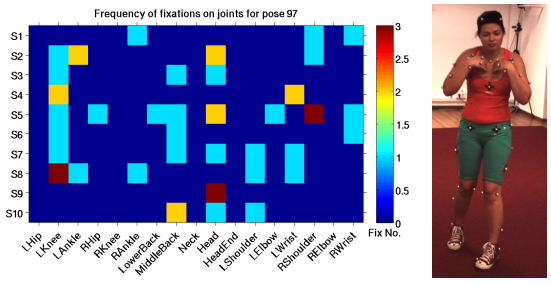


Figure 9. Number of fixations made by each subject, on the 17 body joints, when presented the pose-image shown on the right.

3.2. 3D Pose Re-Enactment

In this section we complement eye-movement studies with an analysis of how well humans are able to reproduce the 3D poses of people shown in images.

What is the joint angle distribution of the poses in our dataset? In fig. 12 we show the angle distribution, for each

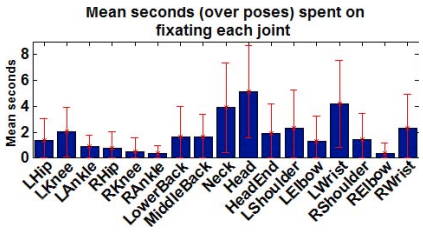


Figure 10. Time spent on fixating each joint. The mean and standard deviation is computed among the 120 poses by aggregating the duration of fixations for all 10 subjects, for each pose.

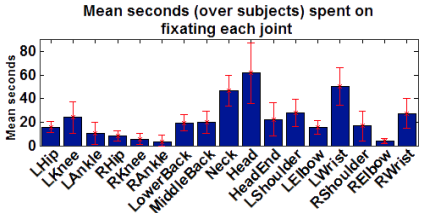


Figure 11. Time spent on fixating each joint. The mean and standard deviation is computed among the 10 subjects by aggregating, for each, the fixation duration over all 120 poses.

joint, measured over the 120 poses in the dataset. Easy poses contain mainly standing positions (very few angles over 30^o in the lower body part), whereas the hard ones, often very different from standing, are spread across all joints.

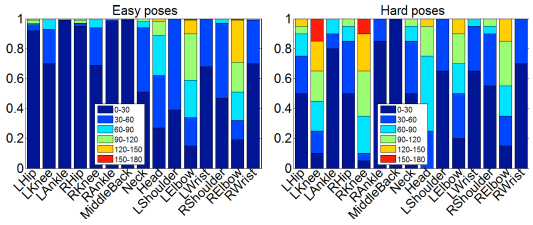


Figure 12. Distribution of joint angles in our dataset (under MPJAE) split over easy (left) and hard poses (right).

Since subjects were asked to match the right and left components of a pose accordingly, we want to know whether there is a balanced distribution between the deviations of our selected poses from a resting pose, over the right and left sides. Fig. 13. shows the mean deviation from a rest pose, over each joint. The angle difference was obtained using MPJAE. There is a similar degree of displacement required for both left and right body sides in replicating poses during experiments.

How long does it take to assume a pose? While there is variance between subjects in the time taken to re-enact a pose, table 1 shows that all of them are consistent in taking more time for hard poses compared to easy ones.

Are easy poses really easy and hard poses really hard? The selection criterion was based on our perception of how

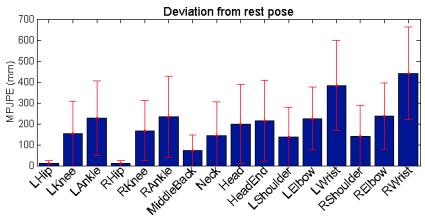


Figure 13. Deviation statistics from rest pose under MPJPE.

Subjects	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Time(sec) Easy	6.6	6.0	4.4	4.4	5.2	6.7	9.2	7.8	4.6	4.8
Time(sec) Hard	9.6	8.2	6.4	8.6	7.5	10.0	11.1	11.5	6.5	7.0

Table 1. Average time for re-enacting a 3D pose. The mean time is 6 ± 1.6s for an easy pose and 8.6 ± 1.8s for a hard one.

hard it would be to re-enact the pose. Here we check how this relates to the measured errors for different poses. The leftmost plot from fig. 14 shows that MPJAE smoothly decreases over time. Subjects require different times to completion for an easy pose. The center pose seems to be perceived as slightly harder, with higher errors and longer completion times. The last image shows a hard pose as well as the errors and time taken to re-enact it. The errors are considerably higher than for the easy poses indicating that our selection of difficult poses indeed resulted in higher re-enactment errors and longer completion times.

How accurately do humans re-enact 3D poses? The subjects decide when they consider completion: body configuration closest to the one shown. Using our error measures we analyze whether their perceived minimum error was indeed the closest they were able to achieve. In table 2 we show that, on average, subject completion errors are worse (by $14 \pm 3\%$ under MPJPE or $9 \pm 10\%$ under MPJAE) than their minimum error achieved during the process of pose re-enactment. The 20 poses we perceived (and selected) as hard to re-enact indeed have larger errors than easy poses by $73 \pm 20\%$ (MPJPE) or $53 \pm 6\%$ (MPJAE). Errors for the different poses are shown in fig. 16.

In a second experiment, not covered here due to space constraints, but described in detail in [13] we allow subjects to adjust their pose while the image is available, thus ruling out short-term memory decay as a confounding factor. We first presented the image only for 5s, removed it, and asked the subjects to re-enact the pose (as in §2.1). Upon completion, we projected the image again, allowing the subjects to adjust their pose once more. For 5 subjects, shown 100 easy and 50 hard poses, the errors were 103.92mm (MPJPE) or 26.32° (MPJAE) (without feedback) and 99.36 mm (MPJPE) or 26.53° (MPJAE) (with visual feedback). The small differences among the two results indicate that continuously available visual stimuli did not significantly change the re-enactment error on completion.

Are there correlations between errors of different body joints? We expect that when subjects misinterpret the posi-

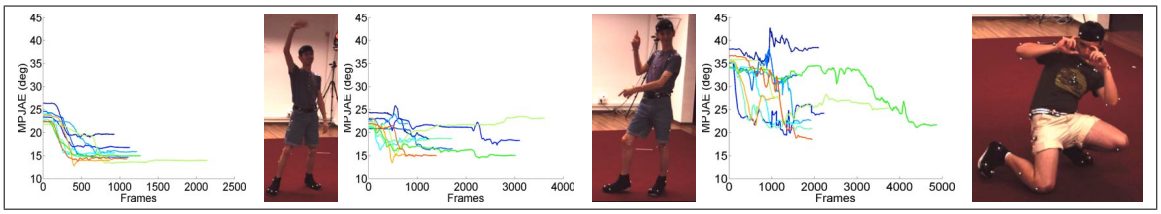


Figure 14. Error variation, over time, for two easy poses (left and center) and a hard one (right).

	MPJPE min error (mm)			MPJPE end error (mm)			MPJAE min error (deg)			MPJAE end error (deg)		
Subjects	Easy	Hard	Both	Easy	Hard	Both	Easy	Hard	Both	Easy	Hard	Both
S1	105.2 ± 34.8	155.0 ± 59.8	113.5 ± 43.9	119.7 ± 38.9	171.1 ± 70.7	128.3 ± 49.3	18.3 ± 6.5	26.4 ± 6.7	19.6 ± 7.2	20.0 ± 6.5	31.0 ± 8.6	21.8 ± 8.0
S2	75.2 ± 24.4	138.2 ± 49.8	85.7 ± 38.0	87.8 ± 27.6	156.1 ± 65.5	99.1 ± 44.4	16.2 ± 5.8	22.9 ± 6.6	17.3 ± 6.4	17.2 ± 6.3	25.3 ± 8.1	18.6 ± 7.2
S3	79.9 ± 34.5	130.0 ± 51.5	88.3 ± 42.0	88.5 ± 37.8	138.3 ± 56.1	96.8 ± 45.1	15.9 ± 5.8	23.5 ± 7.7	17.2 ± 6.8	16.6 ± 5.9	26.7 ± 9.6	18.3 ± 7.6
S4	78.4 ± 32.2	140.0 ± 41.3	88.7 ± 40.8	90.6 ± 36.4	162.7 ± 47.5	102.6 ± 46.8	16.4 ± 5.9	24.0 ± 6.5	17.7 ± 6.6	17.8 ± 6.5	27.2 ± 6.9	19.4 ± 7.4
S5	73.9 ± 28.5	130.2 ± 40.8	83.3 ± 37.2	85.3 ± 29.0	162.2 ± 74.2	98.1 ± 49.1	16.1 ± 5.4	23.4 ± 6.1	17.3 ± 6.2	17.5 ± 5.7	25.8 ± 7.5	18.9 ± 6.7
S6	81.0 ± 38.5	143.7 ± 46.4	91.4 ± 46.1	92.1 ± 43.6	155.3 ± 44.3	102.6 ± 49.6	16.4 ± 6.2	24.4 ± 7.1	17.8 ± 7.0	17.2 ± 6.5	26.7 ± 9.4	18.8 ± 7.9
S7	84.4 ± 33.5	125.3 ± 39.7	91.2 ± 37.7	99.5 ± 39.8	142.2 ± 56.3	106.6 ± 45.6	17.1 ± 6.0	25.4 ± 7.8	18.5 ± 7.0	18.8 ± 6.5	28.0 ± 8.4	20.3 ± 7.7
S8	77.3 ± 25.5	139.2 ± 41.5	87.6 ± 36.8	85.8 ± 29.2	152.4 ± 45.7	96.9 ± 40.8	15.4 ± 6.0	24.6 ± 7.4	17.0 ± 7.1	16.2 ± 6.1	25.9 ± 7.6	17.8 ± 7.3
S9	73.7 ± 30.0	152.0 ± 56.9	86.7 ± 46.1	87.1 ± 32.8	175.2 ± 69.1	101.7 ± 52.4	15.7 ± 5.7	22.4 ± 5.8	16.8 ± 6.2	17.2 ± 5.9	25.2 ± 8.1	18.6 ± 7.0
S10	72.0 ± 23.5	133.9 ± 39.3	82.3 ± 35.2	80.6 ± 25.2	151.9 ± 43.1	92.5 ± 39.2	15.4 ± 5.6	24.3 ± 7.3	16.9 ± 6.7	16.7 ± 5.8	26.5 ± 8.0	18.3 ± 7.2
All	80.1 ± 32.1	138.8 ± 47.0	89.9 ± 41.3	91.7 ± 35.9	156.7 ± 58.1	102.5 ± 47.2	16.3 ± 5.9	24.1 ± 6.9	17.6 ± 6.8	17.5 ± 6.2	26.8 ± 8.2	19.1 ± 7.4

Table 2. Results detailed for easy poses, hard poses as well as over all poses under MPJPE and MPJAE metrics. We display the mean of the minimum errors attained by subjects during re-enactment, as well as the completion errors for the subjects.

tion of a joint, thus exhibiting a large error in that particular articulation, there could be other joints that are wrongly positioned, perhaps to compensate. Fig. 15 indeed shows strong error correlations for the upper body when (under MPJPE) as well as for both arms (under MPJAE).

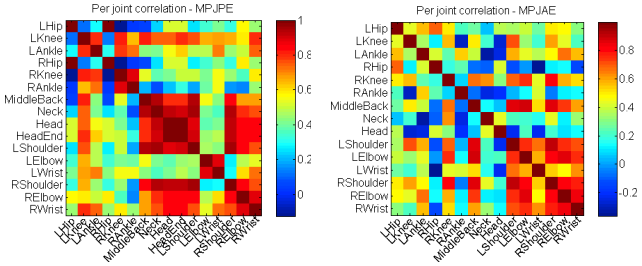


Figure 15. Joint error correlations under MPJPE and MPJAE.

4. Discussion

Our study reveals that people are not significantly better at re-enacting 3D poses given visual stimuli, *on average*, than existing computer vision algorithms[9], at least within the laboratory setup of our study (naturally the errors of computer vision algorithms could be radically different and are subject to our ongoing research). Errors in the order of 10-20° or 100mm per joint are not uncommon. Hard poses selected in the construction of the dataset indeed lead to higher errors compared to easy poses. This indicates that people are not necessarily good at accurate 3D pose recovery, under conventional metrics, a finding consistent with earlier computational studies of 3D monocular human pose ambiguities[21, 22, 19]. Instead, qualitative representations may be used for most tasks, although the implications in skill games (*e.g.* using Kinect[23]) where player's accuracy is valued, but may not be realizable, could be relevant. In the process of reproducing the pose, subjects attend certain joints more than others and this depends on the pose, but the scanpaths are remarkably stable across subjects both spatially and sequentially. Extremities including the head or the wrists are fixated more than internal joints, perhaps because once 'end-effector' positions are known, more constrains are applicable to 'fill-in' intermediate joints on the kinematic chain. Familiar pose sub-configurations are often fixated less (or not at all) compared to unfamiliar ones indicating that a degree of familiar sub-component pose recognition occurs from low-resolution stimuli, not ruling out poselet approaches[4]. An interesting avenue not pursued in almost any artificial recognition system, but not inconsistent with our findings, would be the combination of low-resolution (currently pervasive computer vision) inference with pose and image dependent search strategies that focus on high resolution features–combining bottom-up and selective top-down processing[20, 17, 2].

5. Conclusions

The visual analysis of humans is an active computer vision research area–yet it faces open problems on what elements of meaning should we detect, what elements of the pose should we represent and how, and what are the acceptable levels of accuracy for different human sensing tasks. In this paper we have taken an experimental approach to such questions by investigating the *human pictorial space*, linking images of people with the process in which humans perceive and re-enact those 3D poses. We have developed a novel apparatus for this task, constructed a publicly available dataset, and performed quantitative analysis to reveal

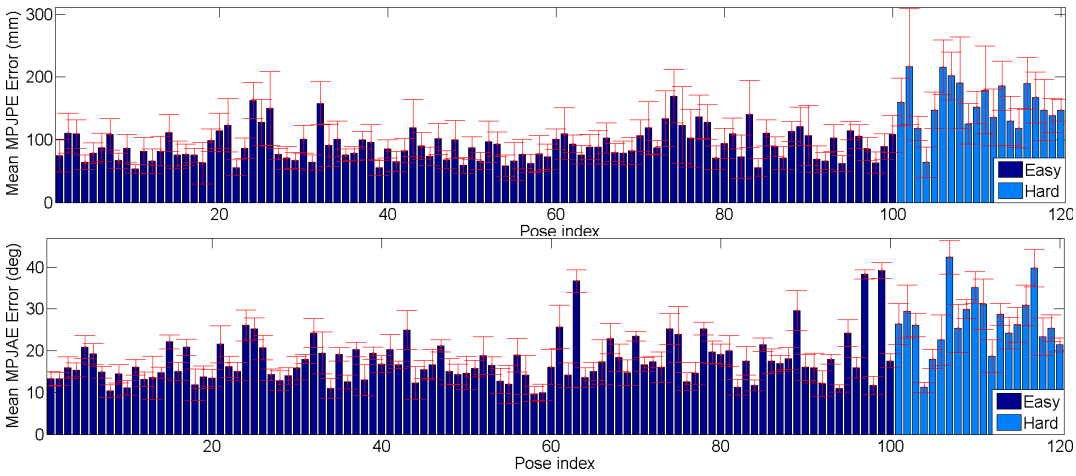


Figure 16. MPJPE and MPJAE versus pose index. Notice significant subject variance and larger errors for 'hard' poses.

the level of human performance, the accuracy in pose re-enactment tasks, as well as the structure of eye movement patterns and the correlation with the pose difficulty. We have also discussed the implications of such findings for the construction of computer-vision based human sensing systems. In future work we plan to design perceptual metrics, as well as person detectors and image search strategies based on our findings and data.

Acknowledgments: Support in part by CNCS-UEFICSDI under CT-ERC-2012-1, PCE-2011-3-0438.

References

[1] A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. *PAMI*, 2006.

[2] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D pose estimation and tracking by detection. In *CVPR*, 2010.

[3] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. *IJCV*, 2010.

[4] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *(ECCV)*, 2010.

[5] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *CVPR*, 2000.

[6] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, 2009.

[7] V. Ferrari, M. Marin, and A. Zisserman. Pose Search: retrieving people using their pose. In *CVPR*, 2009.

[8] J. Gall, B. Rosenhahn, T. Brox, and H. Seidel. Optimization and filtering for human motion capture: A multi-layer framework. *IJCV*, 2010.

[9] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 2014.

[10] G. Johannson. Visual perception of biological motion and a model for its analysis. In *Perception and Psychophysics*, 1973.

[11] J. Koenderink. Pictorial relief. In *Phil. Trans. R. Soc. London. A*, volume 356, 1998.

[12] H. J. Lee and Z. Chen. Determination of 3D human body postures from a single view. *CVGIP*, 30, 1985.

[13] E. Marinoiu, D. Papava, and C. Sminchisescu. Pictorial human spaces: A study on the human perception of 3D articulated poses . Technical report, IMAR and Lund University, 2013.

[14] S. Mathe and C. Sminchisescu. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In *NIPS*, 2013.

[15] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010.

[16] A. Sekunova, M. Black, L. Parkinson, and J. Barton. Viewpoint and pose in body-form adaptation. *Perception*, 2013.

[17] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*, 2007.

[18] L. Sigal, D. J. Fleet, N. F. Troje, and M. Livne. Human attributes from 3D pose tracking. In *ECCV*, 2010.

[19] C. Sminchisescu and A. Jepson. Variational mixture smoothing for non-linear dynamical systems. In *CVPR*, volume 2, pages 608–615, Washington D.C., 2004.

[20] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Learning joint top-down and bottom-up processes for 3D Visual inference. In *CVPR*, 2006.

[21] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3D human tracking. In *CVPR*, 2003.

[22] C. Sminchisescu and B. Triggs. Mapping minima and transitions in visual models. *IJCV*, 61(1), 2005.

[23] M. Sun, P. Kohli, and J. Shotton. Conditional regression forests for human pose estimation. In *CVPR*, 2012.

[24] R. Urtasun, D. Fleet, A. Hertzmann, and P. Fua. Priors for people tracking in small training sets. In *ICCV*, 2005.

[25] D. M. Wolpert, J. Diedrichsen, and J. R. Flanagan. Principles of sensorimotor learning. *Nat. Rev. Neuroscience*, 2011.

[26] Y. Yang and D. Ramanan. Articulated pose estimation using flexible mixture of parts. In *CVPR*, 2011.