Evaluating Example-based Pose Estimation: Experiments on the HumanEva Sets

Ronald Poppe
Human Media Interaction Group, Department of Computer Science
University of Twente, Enschede, The Netherlands
poppe@ewi.utwente.nl

Abstract

We present an example-based approach to pose recovery, using histograms of oriented gradients as image descriptors. Tests on the HumanEva-I and HumanEva-II data sets provide us insight into the strengths and limitations of an example-based approach. We report mean relative 3D errors of approximately 65 mm per joint on HumanEva-I, and 170 mm on HumanEva-II. We discuss our results using single and multiple views. Also, we perform experiments to assess the algorithm's generalization to unseen subjects, actions and viewpoints. We plan to incorporate the temporal aspect of human motion analysis to reduce orientation ambiguities, and increase the pose recovery accuracy.

1. Introduction

Approaches to vision-based human motion analysis can broadly be divided into generative and discriminative. The first category explicitly uses a human body model that describes both the visual and kinematic properties of the human body. Discriminative (recognition) approaches learn the mapping from image space to pose space directly from carefully selected training data.

Human motion analysis from images is challenging due to variations in human body dimensions, camera viewpoint, type of motion and numerous environmental settings such as lighting. In a generative approach, many of these parameters can be included in the estimation. While this may improve the pose recovery accuracy, it comes at the cost of computational complexity. In discriminative approaches, these variations can either be encoded implicitly (for example variations in lighting and body size dimensions), or explicitly (for example viewpoint). The fact that not all parameters can efficiently be encoded explicitly causes discriminative approaches to perform less accurately compared to generative work. However, discriminative approaches are computationally much less expensive, and can potentially be applied in real-time. Therefore, they are useful for applications in Human-Computer Interaction and surveillance, where speed is an issue, rather than accuracy. A discriminative approach also has the ability to automatically (re)initialize. In this respect, a combined generative and discriminative approach (e.g. [13]) is a promising direction to obtain accurate results within reasonable time.

We describe an example-based approach to pose recovery. Such an approach uses a (fixed) set of example poses, with their corresponding visual appearance. Pose recovery is simply selecting the pose that corresponds to the most visually similar example. Since the examples usually cover the pose space very sparsely, the pose estimate is often obtained through interpolation of the n closest examples. While an example-based approach is conceptually hardly complex (in fact, the only parameter that can be tuned is n), existing literature explores many different ways to encode the image and the poses. Moreover, the data used for training and testing is often proprietary, which makes interpretation of the various results and proper consolidation of the findings difficult. The HumanEva data set [12] provides us with the opportunity to compare pose recovery results objectively on data that, although recorded in a controlled environment, poses challenges with respect to the activities performed, and the subjects performing them. This allows us to measure the influence of various factors on the pose recovery performance. In the research described here, we do exactly that, and report our findings on an example-based approach.

The next section discusses related work on example-based pose estimation. Section 3 explains our method. Experiments and our results are presented in Section 4. In Section 5, we summarize the advantages and limitations of our approach, and present directions for future research.

2. Related work

The area of vision-based human motion analysis is too broad to be thoroughly reviewed here. The interested reader is referred to [9] for a recent overview. In this section, we focus on discriminative work, which can be divided into functional and example-based (nearest-neighbor). In functional work, the mapping from image space to pose space

is approached functionally, whereas in example-based work all training data is retained. This notion is important since it puts a practical limit on the number of examples m in the training set of an example-based approach. Regarding the computational complexity, the order of a straightforward example-based algorithm is linear in m. Below, we present a short overview of the different factors that play a role in an example-based pose recovery approach. Subsequently, we will explain how these factors have determined the design of our contribution.

The basis of an example-based approach is to encode the image observations. In existing literature, a number of different image descriptors have been used, usually based on silhouettes [3, 4, 6] or edges [1, 8, 11, 13]. These descriptors vary in terms of computational complexity, descriptor size, robustness against clutter and invariance to scale, rotation and translation. Another important factor is the number of cameras that is employed. Grauman *et al*. [4] show that pose recovery accuracy increases with the number of views. Also, the type of input affects the algorithm's performance. Some authors use a training set of synthesized images [8, 11], which can be encoded robustly without the presence of noise. Others place their synthesized characters against a photographed background to introduce realistic noise into their training set [1, 13]. Closely related are the issues of localization and segmentation. Usually, the location of the person is known, or obtained from a background segmentation process. In the latter case, the success of an algorithm depends heavily on the results of the segmentation.

Human motion analysis is a spatio-temporal problem, hence the temporal aspect can be used to improve accuracy. Howe [6] recovers poses of an entire sequence (batch). This ensures consistency over time, which is particularly useful when only a single view is used. Ong *et al*. use a tracking (incremental) approach, where a dynamical model puts a prior on the poses in the next time frame. This guarantees temporal consistency, and reduces the number of evaluations needed since only poses with a non-negligible prior have to be evaluated. Other measures to reduce the order of example-based algorithms to sub-linear include hashing [11]. Also, a hierarchical organization of the examples can reduce the complexity of the algorithm.

In our contribution, we do not regard the temporal aspect, nor do we apply any measures to reduce the computational complexity. A variant of histogram of oriented gradients (HoG, [2]) is used to encode the image observations within extracted foreground masks. We vary the number of views and report the performance on the majority of the available actions and subjects in both the HumanEva-I and HumanEva-II test sets. Our results provide insight into the capabilities and limitations of example-based approaches.

3. Method

In this section, we will describe the components of our approach. Section 3.1 discusses the image descriptors that are used in this research. The pose estimation approach is explained in Section 3.2.

3.1. Histogram of oriented gradients

Dalal and Triggs proposed histograms of oriented gradients (HoG) as an image descriptor to localize pedestrians in cluttered images [2]. We believe that their use can be extended to pose recovery. Gradients are to some extent invariant to lighting changes. Moreover, spatial ordering is preserved, which has been found to be of key importance for effective recovery [10].

HoGs are calculated within an image's region of interest (ROI), in our case the bounding box around the subject. While HoGs can be used to determine this region, as in [2], we rely on background subtraction. This significantly speeds up the process, and we obtain the foreground mask at the same time. We describe the process here in detail to allow for replication. First, we apply the background subtraction with the suggested risk values, as included in the HumanEva source code [12]. The minimum enclosing box of all foreground areas larger than 600 pixels is obtained. After conversion to HSV color space, we apply shadow removal in the lower 20% of the ROI. Pixels that have a saturation that is between 0 and 25 higher than the saturation of any of the means in the background mixture model, are removed from the foreground mask. We again obtain the minimum enclosing box, which is our ROI. Figure 1 shows an example of background subtraction and shadow removal.

(a) (b) (c)

Figure 1. Example of foreground mask calculation: (a) original image, (b) with background subtracted, and (c) shadow removed.

It may seem that our approach is highly sensitive to good background subtraction, but the shadow removal is only needed to ensure that the ROI fits the subject reasonably.

For certain cases, we slightly adjusted the parameters. For camera 1 in HumanEva-I, only for subject 2, we multiplied the risk with factor 10^{12} to remove artefacts from the foreground. For cameras 2 and 3 in HumanEva-I, we lowered the shadow threshold to 10. We did not use the additional four grayscale cameras in HumanEva-I. For HumanEva-II, we only reduced the background risk with factor 10^{50} for camera 3.

To obtain the HoG, we divide the ROI into a grid with 6 rows and 5 columns. This is an arbitrary choice, but the height of each cell roughly corresponds with the height of the head in a standing position. Similarly, in a relaxed standing pose, the body covers approximately 3 columns horizontally. We did not perform experiments with various settings but admit this would have been interesting.

Within each cell in the grid, we calculate the orientation and magnitude of each pixel that appears in the foreground mask. We divide the absolute orientations over 9 equally sized bins in the 0°- 180° range. Each pixel contributes the magnitude of its orientation to the according histogram bin, which results in a 9-bin histogram per cell. The total length of the descriptor is therefore 270. The entire descriptor is normalized to unit length. Compared to Dalal and Triggs [2], we do not apply color normalization and do not use the notion of blocks (groups of cells). This reduces computation cost, and results in a significantly reduced descriptor size.

3.2. Pose estimation

In an example-based approach, each image observation is encoded and matched against a training set of encoded observations. We use the previously described HoGs as encodings. To match a HoG with those in the training set, we need to define a distance measure between the two descriptors. We performed a small-scale experiment with Manhattan, Euclidian, cosine and χ^2 distance. Consistent with earlier findings, Manhattan and Euclidian distance proved to be most suitable. In this work, we use the Manhattan distance since it has a lower computational complexity.

Matching a HoG with the entire training set results in a distance value for each of the m examples. We could choose the example with the lowest distance, as this is the example that best matches the image observation. However, in practice, taking the n best matches results in more accurate pose recovery. Of course, n will depend on the number of examples in the training set that are close to the presented frame. Here we use $n = 25$, in accordance with [10]. To determine the final pose estimate, we use the poses that correspond to the n best examples. The final estimate is the normalized weighted interpolation of these poses. This implies that close HoG matches, contribute more to the final estimate than HoG matches at a larger distance. One word of caution is in its place here. Since we interpolate poses, the final joint estimates are likely to lie closer to the mean distance for this joint, so closer to the body. This effect is especially visible for examples that have similar image observations but are distant in pose space.

Since we do not determine the correspondences between our localized subject in the image and the estimated pose, we are not able to reliably estimate the global position of each joint. Instead, we report the distances of each joint relative to the pelvis (torsoDistal) joint.

4. Experimental results

We report our findings of experiments on both the HumanEva-I and HumanEva-II data set. First we describe our two training sets, one for monocular pose recovery (T1), and one for recovery using three cameras simultaneously (T3). In both cases, we used the HumanEva-I sequences that are used for training and validation.

4.1. Training sets

For the monocular training set T1, we associate the HoGs for each view with their corresponding poses. Only the examples that contain valid mocap data are included in the training set.

	Subject			
Action	S1	S2	S3	Total
Walking	1176	876	895	1899
Jog	439	795	831	1696
Throw/Catch	217	806	0	2065
Gestures	801	681	214	1023
Box	502	933	933	2947
Combo	0	0	0	0
Total	3135	3622	2873	9630

Table 1. Number of training examples per action and subject.

When given a new image observation, together with the knowledge from which camera the observation is obtained, we can now estimate the pose. We observe that the elevation (rotation in vertical direction) and roll (rotation around line of sight) of all cameras are approximately the same. In other words, the orientation of all cameras is almost equal except for the orientation around a vertical axis. If we would rotate the subject in the scene around a vertical axis, we would theoretically be able to generate very similar observations for all cameras. In practice, view-specific parameters such as backgrounds and lighting conditions are likely to result in observations that are somewhat different. However, we want our approach to be robust against these image deformations and therefore, we perform this rotation virtually. This has the additional advantage that the number of training data is effectively tripled, resulting in a total of 28,890 samples.

We transform the mocap data in such a way that we obtain the global positions as if we were looking through another camera. With an observation from camera i, and the projection onto camera j, our pose vector $\mathbf{p_i} = (x_i, y_i, z_i, 1)^{-1}$ is transformed into $\mathbf{p_j}$ as follows: $\mathbf{p_j} = M_j M_i^{-1} \mathbf{p_i}$, where M_i and M_j are the rotation matrices of cameras i and j, respectively.

In T3, we combine the HoGs of the three views into a single HoG descriptor of length 810. This combined descriptor is larger, and contains the same pose from multiple views. We therefore expect increased pose estimation accuracy over the monocular descriptors. However, combining our HoGs comes at a cost of a reduced number of examples, compared to T1. For each frame with valid mocap, we have exactly one example. The total amount of examples m that we can obtain is 9,630. Table 1 summarizes the origins of the examples.

Combining all views into one descriptor has some drawbacks. When the setup of the cameras changes, the descriptor cannot be used anymore. The relative orientations of the cameras are encoded implicitly in the combined descriptors. Practically, this means that we cannot evaluate the HumanEva-II sequences with our training sets, since these are obtained from the HumanEva-I setting. Another drawback arises when one or more views contain inaccurate segmentations. This can, in some cases, render the example useless.

4.2. Results for HumanEva-I

The HumanEva-I test sequences are performed with the same camera setup as the training sets. Also, the same test subjects appear in these videos. Except for action Combo, all action-subject combinations also exist in the training set. For the number of examples per subject and action, we again refer to Table 1. We did not perform evaluations for subject 4, due to background segmentation errors. We plan to report on these sequences in the near future, to provide more insight into how our approach performs on unseen subjects.

For each sequence, we evaluate the performance on both T1 and T3. For T1, we use the image observations from camera 1. In informal experiments we looked at the differences between the three cameras but these were small. For T3, we use the combined image observations of all three cameras. The results, broken down by action, subject and training set, are summarized in Table 2. We omitted all frames for which the mocap data was invalid. Also, the first 5 frames of each sequence were removed since these frames were duplicated in the decoding of the video.

The first thing to notice is the relatively small difference between the performance of our monocular tests, and those using three cameras. We obtained mean errors over all subjects and sequences of 66.22 and 60.93 *mm*, respectively. Apparently, our HoG descriptors are to a large extent invariant to depth ambiguities. We expect that this ability can mainly be subscribed to the fact that our descriptor is based on gradients, which have high magnitudes for the arms. These are helpful when differentiating between front and back poses.

To be able to interpret our results, we need a baseline. We use the mean distance between all poses. Since we interpolate the poses of the best matches, our pose estimate will always be within the range of poses in the training set. This baseline is therefore the mean distance for randomly selecting a pose with $n = 1$. The mean distance per joint for T3 is then 299.87 *mm*. For T1, the distance is slightly lower, 291.54 *mm*.

4.2.1 Comparison with related research

Before we take a closer look at the errors for different sequences, we compare our general findings with those previously reported on the HumanEva-I data set. We restrict our comparisons to discriminative approaches.

The work by Howe [5] uses silhouettes in an example-based approach. Attention is paid to correct segmentation of these silhouettes. As in previous work [6], chamfer distance and turning angle are used to compare the silhouettes. Additionally, optical flow is used to include some form of temporal consistency. Second order Markov chaining is used to eliminate poses that have matching observations, but differ significantly in pose (e.g. swapping of the arms). The fact that silhouettes are ambiguous regarding the depth ordering of limbs makes such a measure necessary. Errors are reported in 2D for the Walking sequence of subject 3. Tests are performed with either one color camera, or one grayscale camera, with comparable results. The mean error in both cases is around 17 pixels which, for a figure height between 250 and 410 pixels and an assumed subject height of 1.80 meters, comes down to approximately 75-125 *mm*. If we compare these results with our own, there are some important differences. First, Howe performs image registration and the errors reported are in fact absolute. Second, our approach recovers poses in 3D and does not make use of temporal information. We would be interested to see how our approach would perform with the silhouettes and matching functions as used by Howe. This would give more insight into the strengths and weaknesses of both image descriptors.

Lee and Elgammal [7] present another study that reports on the HumanEva-I set. Silhouettes are used, and a functional mapping is learned. In an offline step, they construct a manifold topology based on synthesized silhouettes of a walking person, as viewed from a large number of angles. The novelty of their work is to learn a geometric mapping between the embedded manifold to the data. The authors re-

	Subject 1		Subject 2		Subject 3	
Action	C1	C1–3	C1	C1–3	C1	C1–3
Walking	41.24 (16.84)	37.54 (11.95)	39.56 (26.75)	40.09 (23.86)	55.27 (21.70)	55.25 (25.07)
Jog	46.38 (18.60)	45.21 (13.74)	38.02 (9.97)	37.65 (12.20)	47.35 (23.47)	45.37 (18.32)
Throw/Catch			69.49 (31.94)	57.61 (23.75)	111.71 (33.38)	92.79 (31.75)
Gestures	26.38 (13.51)	23.69 (7.18)	75.13 (28.10)	72.83 (26.32)	75.29 (11.09)	56.11 (6.44)
Box	79.71 (27.71)	88.67 (36.20)	103.37 (45.08)	91.28 (41.19)	100.35 (52.31)	92.28 (47.86)
Combo			69.84 (49.34)	71.89 (52.17)	106.11 (79.74)	83.89 (53.10)
Average	48.32 (19.17)	48.78 (17.27)	65.90 (31.86)	61.89 (29.92)	82.68 (39.95)	70.95 (30.42)

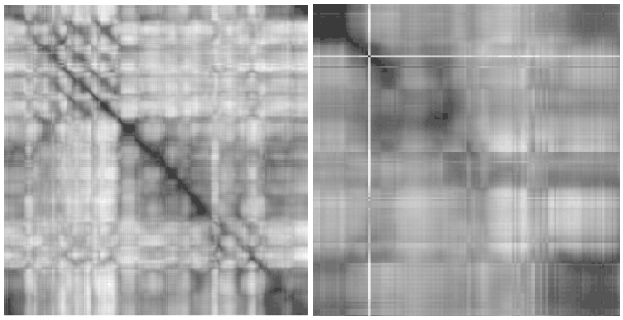
Table 2. Mean 3D error (and SD) in *mm* for HumanEva-I test sequences, evaluated with a single camera (C1) and all three cameras (C1–C3).

port errors on the Walking validation sequences of subjects 1-3. Their mean, normalized, relative 3D error is approximately 35 *mm* when no dynamical model is assumed, and 31 *mm* using a particle filter with constant velocity for both gait phase and view change. Normalization of the pose is obtained by transforming the pose coordinates into a body-centered coordinate space. Effectively, this removes errors due to incorrect estimation of the pose around the vertical axis. Insensitivity to the subject is demonstrated since the manifold topology is learned from data not contained within HumanEva-I. It remains to be investigated how the approach works for open (i.e. non-repetitive) actions.

4.2.2 Results for individual sequences

We will now discuss our results in more detail. If we look at different actions, we see large variations between sequences. In general, poses from the Walking and Jog sequences are recovered with the highest accuracy. This can, at least partly, be explained by the fact that these motions have been performed at least several times in the training sets. Each cycle resembles the others in contrast to, for example, catching a ball where the ball appears at more or less random places. In Figure 2, we present the affinity matrices of two walking cycles and two instances of catching and throwing a ball. We immediately see the similarity between the two walking cycles by the dark line on the diagonal. This line is less apparent for two sequences of catching and throwing a ball.

We should be able to see the same when looking at our test results. Figure 3(a-b) shows the plots of the mean errors over the sequences Walking and Throw/Catch, performed by subject 2 and obtained using all three cameras. The error plot is much more peaked for the Throw/Catch sequence. The peaks (e.g. around 350, 550 and 700) correspond to catching or throwing the ball. The lower errors around 450 and 650 correspond to waiting for the ball, which is in fact a standing pose. The slightly higher errors in the Walking sequence after 200, and around 650, correspond to the subject walking towards the camera. Here, depth ambiguities occur.



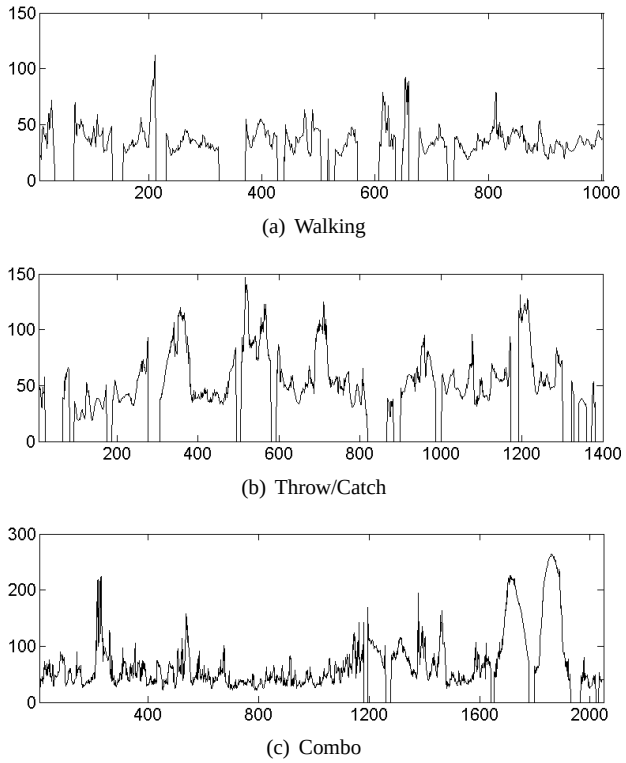
(a) Walking (b) Throw/Catch

Figure 2. Affinity matrices for camera 1 of (a) walking cycle, and (b) Throw/Catch sequence, performed by subject 1. Dark values correspond with small distances. Notice the similarity between the walking cycles. The white 'plus' in the Throw/Catch sequence is caused by incorrect segmentation due to the presence of the ball.

When we take another look at Table 2, we notice rather larger differences between subjects for the Gesture action. In this action, the subject waves and makes beckoning gestures. Subject 1 and 3 perform these gestures with their right hand, in both the training and test sequences. This explains the low standard deviations.

The Box actions show some of the highest errors, which is somewhat surprising. More careful analysis of the video data shows that, for subject 2 and 3, there is quite some variation in the footwork. Subject 1 uses the same standing pose as a basis but is facing camera 1. From this view, it is difficult to estimate the depth of the arms. The view for camera 2 and 3 are almost exactly from the side. This results in many estimations where the wrong arm is estimated to be stretched out.

The Combo action is a combination of walking, jogging, and some 'freestyle' moves, jumping on one leg and balancing on one foot. These latter moves are not present in the training set. For subject 2, the mean error over the whole sequence is shown in Figure 3(c). The peak around frame 250 is caused by incorrect background segmentation. Frames 1160-1370 contain the jumping on one leg action, in frames



(a) Walking

(b) Throw/Catch

(c) Combo

Figure 3. Mean 3D error (in *mm*) plots for HumanEva-I (a) Walking, (b) Throw/Catch, and (c) Combo action, all performed by subject 2 and obtained using all three cameras. Instances that have a zero error contain invalid mocap.

1660-1960 the subject balances on one foot. The difference in error is apparent, and gives us a clue about the performance of an example-based approach for unseen actions. We discuss this further in Section 4.3.

4.2.3 Additional tests on validation set

To allow comparison with previous work on the HumanEva-I set, we report here our results for the Walking sequence, performed by subject 1. This sequence is contained within our training sets, so we had to remove it. Our resulting training sets with removed trial are $T1_T$ and $T3_T$, for the monocular and multi-camera cases, respectively. In addition, we created training sets where we removed all examples for the Walking action, $T1_A$ and $T3_A$. This reduces the training set by 20%. Also, we removed all instances of subject 1, resulting in $T1_S$ and $T3_S$, each containing roughly two thirds of the total number of samples. These training sets allow us to gain insight into the generalization to unseen subjects and unseen motions. The results are presented in Table 3, with mean errors for the training part (frames 591-1203), the validation part (1-590) and over the entire sequence.

Compared to the results in Table 2, the errors obtained

Set	Train	Validation	Total
$T1_T$	80.14 (25.39)	74.40 (23.95)	77.15 (24.81)
$T3_T$	76.92 (28.50)	74.81 (23.86)	75.82 (26.20)
$T1_A$	94.72 (33.46)	92.69 (28.90)	93.66 (31.17)
$T3_A$	103.90 (46.23)	110.55 (44.96)	107.37 (45.68)
$T1_S$	84.18 (29.68)	76.86 (25.56)	80.36 (27.83)
$T3_S$	78.25 (29.02)	77.30 (24.12)	77.76 (26.57)

Table 3. Mean 3D error (and SD) in *mm* for HumanEva-I training/validation sequence Walking, performed by subject 1. Errors are presented for different training sets.

here are approximately 35-70 *mm* per joint higher. It should be noted that the number of examples is reduced and this is likely to have a negative effect on the pose recovery accuracy. Our first observation is that even when only one trial is removed, the error is larger. Removal of this sequence results in a lower number of examples of the Walking action. Moreover, all examples of the same subject for the Walking action are left out. Apparently, our approach is much more accurate when person-specific observations are used.

When all Walking examples are removed, the error increases 15-35 *mm*. Closer analysis of the examples used in the reconstruction shows that these are mainly from the Jog action. Walking and jogging show many resemblances, but in the Jog action the elbows are usually more bent, and the distance between the feet is kept small. This explains the increase.

A final observation can be made with respect to the removal of examples containing subject 1. The results are slightly less accurate than in the case where only the Walking action of subject 1 is removed. This leads us to conclude that only few examples from other actions are used in the estimation. Also, we have an estimate of the accuracy for pose recovery when the subject is not in the training set. Again, we must make a remark about the significantly reduced number of relevant samples, that is likely to increase the error also in this case.

4.3. Results for HumanEva-II

The HumanEva-II set consists of two Combo sequences performed by subjects 2 and 4. HumanEva-II differs from HumanEva-I in the recording setup. Four color cameras are used, instead of the three color cameras and four grayscale cameras. The cameras are placed at different positions, and the elevation angles differ slightly. As mentioned before, this setup does not allow us to do any evaluation on the T3 training set.

Our evaluation was performed similar to that of the HumanEva-I test sequences. In Table 4 the results are summarized for both subjects and for each camera separately. The three sets contain various movements within the sequence. Set 1 (frame 1-350) contains walking, set 2 (frame 1-700) contains walking and jogging, and set 3 (frame 1-

1202 for subject 2, and 1-1258 for subject 4) contains the whole sequence. This includes walking, jogging, jumping on one leg and balancing on one foot.

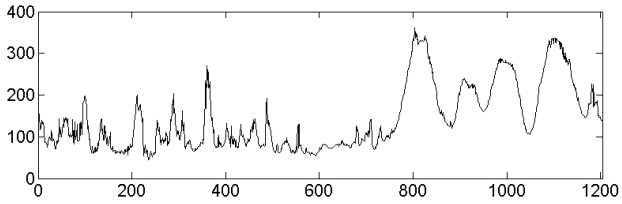
Our first remark concerns the analysis of the sequence performed by subject 4. We removed frames 298-337 from the results since these appeared to be erroneous. We obtained mean errors per joint above 1,200 *mm*. Visual inspection of our pose estimates and the video revealed no peculiarities. Moreover, an average distance per joint of 1,200 *mm* is very unlikely since we use relative distances.

Compared to the results that we obtained for the Combo sequences of HumanEva-I for T1, these results are higher. Ideally, we would expect errors for set 1 that are comparable to those of the Walking sequences in HumanEva-I. Similarly, for set 2, we expected results that are close to those of Walking and Jog.

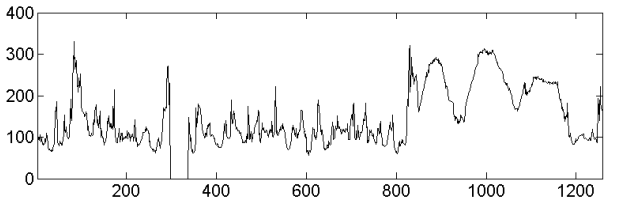
In explaining these results, we take another look at the results that we obtained when analyzing the validation Walking sequence in HumanEva-I, Section 4.2.3. We observed that the error increases when there are no examples in the training set for the same person, performing the same action. We expect that this is also the case for our HumanEva-II results. Although subject 2 appears in HumanEva-I, the clothing is different. This probably has an effect on the HoGs, and subsequently on the closest matches.

Yet, if we consider the subjects as unseen, the error is higher. We expect this to be the result of the modified camera setup. The elevation of the camera is different, and there is a roll angle introduced for some cameras. We expect this to have quite a large impact, as HoG are not rotation invariant. Also, the background is modelled with a single Gaussian. This results in less accurate background segmentation, and this reflects on the HoG descriptors.

Figure 4 shows the mean error plots of the HumanEva-II sequences. The increased error for the unseen actions (frame 750-end) is apparent. Other peaks (e.g. for subject 2 around frame 100, 220, 300 and 370) are due to forward-backward ambiguities. Figure 5 shows the original frames and the frames corresponding to the single best example. We plan to conduct experiments with tracking to increase accuracy in these cases.



(a) Subject 2



(b) Subject 4

Figure 4. Mean 3D error (in *mm*) plots for the HumanEva-II Combo sequences performed by (a) subject 2, and (b) subject 4. Instances that have a zero error have been ignored (see text).

5. Conclusion and future work

We took an example-based approach to pose recovery. Histograms of oriented gradients have been used as image descriptors, and matched against our training sets. The final pose estimate was obtained by interpolating the poses of the 25 best matches based on their normalized weight. Since we do not calculate the correspondence between the image and our examples, we report mean 3D errors over all joints, relative to the pelvis.

We evaluated our approach on both the HumanEva-I and HumanEva-II test sequences. For HumanEva-I, we found only small differences between our monocular case, and when three views were used. Part of this small difference can be explained by the reduced amount of relevant samples in the multi-view case. Our results varied depending on the action performed. For walking and jogging, errors around 45 *mm* were obtained. The mean error for all actions and all subjects is approximately 65 *mm*. Our approach is somewhat person specific, and does not generalize well to unseen actions. For HumanEva-II, errors close to 120 *mm* were obtained for the walking and jogging movements. The higher error can be explained by the fact that the subjects look different from the ones in the training set (one of the subjects does not appear in the training set, the other wears different clothing). Also, the viewpoints and recording settings are slightly different, which resulted in less accurate background segmentation and consequently, less accurate HoG descriptors. Due to this different setup, we were not able to perform any evaluations with more views.

From a practical point of view, we plan to process the sequences containing subject 4 in HumanEva-I. This would give us more insight into whether the reduced accuracy for the HumanEva-II sequences is due to the changed viewpoints, or the fact that the subjects appear differently.

In order to allow our approach to perform in real-time, we would like to evaluate some measures to reduce the computational complexity, including hashing and hierarchical matching.

We have already observed that vision-based human motion analysis is spatio-temporal in nature. We noticed several times that reduced accuracy is often caused by forward-

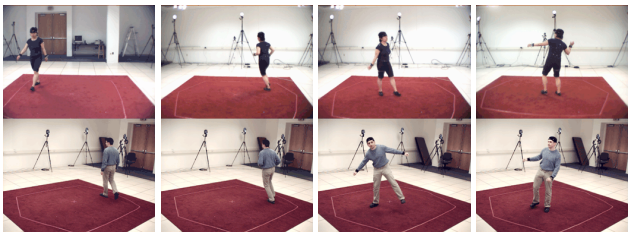
		Subject 1	Subject 2		Subject 3	
Action	Set	Camera 1	Camera 2	Camera 3	Camera 4	Average
Combo S2	1	121.96 (72.51)	100.35 (34.28)	101.25 (44.72)	113.49 (48.19)	109.26 (49.93)
Combo S2	2	111.96 (59.83)	95.93 (34.54)	105.97 (60.93)	116.06 (54.85)	107.48 (52.54)
Combo S2	3	173.92 (111.22)	142.62 (77.93)	203.13 (144.72)	161.94 (89.45)	170.40 (105.83)
Combo S4	1	129.92 (54.38)	120.73 (49.24)	183.61 (86.69)	113.49 (48.19)	136.94 (59.63)
Combo S4	2	138.89 (58.27)	115.93 (39.16)	146.75 (71.99)	116.06 (54.85)	129.41 (56.07)
Combo S4	3	166.31 (77.09)	147.50 (66.10)	200.31 (120.20)	161.94 (89.45)	169.02 (88.21)

Table 4. Mean 3D error (and SD) in *mm* for HumanEva-II test sequences, evaluated with a single camera. Results are broken down per set.

backward ambiguities. We therefore plan to conduct evaluations where we track the movements over time (incrementally), or recover poses in a batch approach over an entire sequence. We expect both approaches to substantially reduce our errors, and bring accurate real-time pose recovery within reach.

Acknowledgements

This work was supported by the European IST Programme Project FP6-033812 (Augmented Multi-party Interaction with Distant Access, publication AMIDA-25), and is part of the ICIS program. ICIS is sponsored by the Dutch government under contract BSIK03024.



(a) Frame 300 (b) Frame 600 (c) Frame 900 (d) Frame 1200

Figure 5. Single best estimate (top row) and original frame (bottom row) for the HumanEva-II Combo sequence performed by subject 2. Frame 300 shows forward-backward ambiguity in the walking action. Frames 900 and 1200 contain unseen movements. Colors have been adapted for the viewer's convenience.

References

[1] A. Agarwal and B. Triggs. A local basis representation for estimating human pose from cluttered images. In *Proceedings of the Asian Conference on Computer Vision (ACCV'06) - part 1*, number 3851 in Lecture Notes in Computer Science, pages 50–59, Hyderabad, India, January 2006. 2

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05) - volume 1*, pages 886–893, San Diego, CA, June 2005. 2, 3

[3] A. Elgammal and C.-S. Lee. Nonlinear manifold learning for dynamic shape and dynamic appearance. *Computer Vision and Image Understanding (CVIU)*, 106(1):31–46, April 2007. 2

[4] K. Grauman, G. Shakhnarovich, and T. Darrell. Inferring 3D structure with a statistical image-based shape model. In *Proceedings of the International Conference on Computer Vision (ICCV'03) - volume 1*, pages 641–647, Nice, France, October 2003. 2

[5] N. R. Howe. Evaluating lookup-based monocular human pose tracking on the humanEva test data. In *Proceedings of the Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHuM)*, Whistler, Canada, December 2006. 4

[6] N. R. Howe. Silhouette lookup for monocular 3D pose tracking. *Image and Vision Computing*, 25(3):331–341, March 2007. 2, 4

[7] C.-S. Lee and A. Elgammal. Body pose tracking from uncalibrated camera using supervised manifold learning. In *Proceedings of the Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHuM)*, Whistler, Canada, December 2006. 4

[8] E.-J. Ong, A. S. Micilotta, R. Bowden, and A. Hilton. Viewpoint invariant exemplar-based 3D human tracking. *Computer Vision and Image Understanding (CVIU)*, 104(2-3):178–189, November 2006. 2

[9] R. Poppe. Vision-based human motion analysis: an overview. *Computer Vision and Image Understanding (CVIU)*, to appear. 1

[10] R. Poppe and M. Poel. Comparison of silhouette shape descriptors for example-based human pose recovery. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'06)*, pages 541–546, Southampton, UK, April 2006. 2, 3

[11] G. Shakhnarovich, P. A. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *Proceedings of the International Conference on Computer Vision (ICCV'03) - volume 2*, pages 750–759, Nice, France, October 2003. 2

[12] L. Sigal and M. J. Black. HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, Department of Computer Science, Providence, RI, September 2006. 1, 2

[13] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Learning joint top-down and bottom-up processes for 3D visual inference. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06) - volume 2*, pages 1743–1752, New York, NY, June 2006. 1, 2