Example-based Pose Estimation in Monocular Images Using Compact Fourier Descriptors*

Ronald Poppe and Mannes Poel
University of Twente, Human Media Interaction Group
P.O. Box 217, 7500 AE Enschede, the Netherlands
{poppe,mpoel}@ewi.utwente.nl

Abstract

Automatically estimating human poses from visual input is useful but challenging due to variations in image space and the high dimensionality of the pose space. In this paper, we assume that a human silhouette can be extracted from monocular visual input. We compare the recovery performance of Fourier descriptors with a number of coefficients between 8 and 128, and two different sampling methods. An example-based approach is taken to recover upper body poses from the descriptors. We test the robustness of our approach by investigating how shape deformations due to changes in body dimensions, viewpoint and noise affect the recovery of the pose. The average error per joint is approximately 16-17° for equidistant sampling and slightly higher for extreme point sampling. Increasing the number of descriptors does not have any influence on the performance. Noise and small changes in viewpoint have only a very small effect on the recovery performance but we obtain higher error scores when recovering poses using silhouettes from a person with different body dimensions.

*This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (FP6-506811, publication AMI-93), and is part of the ICIS program. ICIS is sponsored by the Dutch government under contract BSIK03024.

1 Introduction

Being able to automatically estimate human poses from visual input is useful in many application domains, including surveillance, animation and human-computer interaction (HCI). However, the problem is difficult since the relation between image observations and poses is multivalued in both directions. Variations in human body dimensions, appearance, and environmental settings such as lighting conditions and camera viewpoint possibly result in many observations for the same pose. On the other hand, similar observations can correspond to a range of poses due to projection, (self)occlusions and limited visual accuracy. All these parameters, and the large number of degrees of freedom (DOF) in the human body, inhibit an exhaustive search. Therefore, many approaches adopt a (detailed) human body model that describes how the human body appears in the image space. Poses are estimated by optimizing the error between visual input and the projection of the pose to image space. One problem with these *model-based* approaches is that initialization of both model and pose is often difficult. *Model-free* approaches do not use an explicit human body model but instead learn a mapping from image space to pose space.

Usually, image features are extracted from the visual input to allow more efficient matching. Features that are often used are edges, silhouettes, color and motion. Here we focus on silhouettes because they can be extracted relatively robustly from images; they are insensitive to variations in surface such as color and texture; and they encode a great deal of information to recover 3D poses [Agarwal and Triggs, 2004]. However, performance is limited due to artifacts such as shadow and noisy background segmentation, and it is often difficult or impossible to recover certain degrees of freedom due to the lack of depth information.

In this paper, we assume that a human silhouette can be extracted from monocular visual input. We compare the recovery performance of Fourier descriptors with five different lengths and two different sampling methods. An example-based approach is taken to recover upper body poses from the descriptors. In this research pose estimation reduces to the estimation of joint angles in the upper body. Our output space is a 9-dimensional vector corresponding to 9 DOF in the upper body, as summarized in Table 1. We test the robustness of the Fourier descriptors by investigating how shape deformations due to changes in body dimensions, viewpoint and noise affect the recovery of the pose. Note that we only investigate the robustness of the descriptors without modeling temporal dependencies or including statistical information about pose frequency. These factors can improve estimation. Therefore, the estimation errors that are reported in this paper are higher than if we used this additional information.

The paper is organized as follows. Section 2 summarizes related work on pose estimation. Fourier descriptors and the two sampling methods are discussed in Section 3, followed by the description of our pose recovery approach in Section 4. Experiment setup, results and analysis are discussed in Section 5 and we conclude in Section 6.

2 Related work

Vision-based human pose recovery techniques enable estimation of human poses and movement without obtrusive or expensive equipment. An extensive overview of the topic can be found in Gavrila [1999] and Moeslund and Granum [2001]. Current research can roughly be divided into *model-based* and *learning-based* approaches. Model-based approaches [Delamarre and Faugeras, 1999, Deutscher et al., 2000, Sidenbladh et al., 2000] presuppose an explicitly known parametric body model. The pose recovery problem is typically solved by matching the pose variables to a forward rendered human model based on labelled extracted features. Drawbacks of these approaches are the often difficult labelling and initialization. Estimating the pose has many local minima which can lead to low performance [Sminchisescu and Triggs, 2003]. Learning-based approaches [Howe, 2004, Shakhnarovich et al., 2003, Agarwal and Triggs, 2004] do not assume an explicit human body model. Instead, a relation from extracted features to pose variables is learned from training data. In *example-based* approaches, a subcase of learning-based approaches, a collection of images or image features is stored together with their corresponding pose description. For a given input image, a similarity search is performed and the poses are interpolated. Learning-based approaches are viewpoint dependent and require a large amount of training data, especially when many DOF are modeled or the allowed motion is unconstrained.

We focus on learning-based pose recovery. The key point of these approaches is to have a robust image descriptor. This descriptor should be able to generalize over variations in pose observation but distinguish between different poses. This is a difficult requirement, and many different image descriptors have been evaluated throughout literature. Silhouettes and edges are used the most, because they can be easily extracted and are, to some extent, lighting invariant.

Howe [2004] uses silhouettes which are matched to a collection of known poses using turning angle and Chamfer distance. Agarwal and Triggs [2004] encode the silhouette boundary using shape contexts [Belongie et al., 2002] and use Bayesian non-linear regression to recover 54 DOF body poses with high accuracy. Elgammal and Lee [2004] learn view-based activity manifolds from silhouettes. Mappings from silhouette to activity manifold and from activity manifold to pose are learned from training data. Brand [1999] also learns a mapping from silhouettes to poses but does not use an intermediary level. Pose estimation over an entire sequence is performed by applying Viterbi algorithm to a Hidden Markov Model (HMM). Observations are silhouette central moments, hidden states correspond to the pose parameters. Moments are also used by Rosales and Sclaroff [2000]. They observe that an inverse mapping from image space to pose space cannot be implemented by a single function. Therefore, the pose space is divided into clusters and specialized functions are learned for each cluster from Hu moments [Hu, 1962] to the 2D pose space. A neural network is used as mapping function. In Rosales et al. [2001], the work is extended to allow input from multiple cameras. Another multi-camera approach has been taken by Ren et al. [2004]. They recognize sequences of motions using a motion

database that contains discriminative local features extracted from silhouettes obtained from three cameras. The features are learned from a motion database in a preprocessing step.

Edges contain more information but are also more sensitive to texture. Mori and Malik [2002] extract shape contexts of edge points from an image. They store an example collection to recover the 2D joint positions, that are transformed to a 3D pose estimation in a subsequent step. In a similar approach, Sullivan and Carlsson [2002] store sets of exemplars which they call key frames. A point correspondence algorithm is employed to calculate the distance between the normalized edges of an image and an exemplar, and to recover joint locations. Shakhnarovich et al. [2003] use edge direction histograms within a contour and apply an efficient search mechanism to find corresponding upper body poses from an example set.

Dynamics are often used to improve pose recovery. Deutscher et al. [2000] use a particle filter to propagate movements in time. Another approach is to learn the dynamics of human movement from training samples [Sidenbladh et al., 2000, Agarwal and Triggs, 2004, Elgammal and Lee, 2004]. Although this often leads to more stable and accurate estimation results, it also puts a strong prior on the movements that can be recovered.

3 Silhouette shape descriptors

We only compare shape descriptors, thus ignoring the dynamics. Therefore, our work is also closely related to content based image retrieval. There exist a large number of shape descriptors, see Veltkamp and Hagedoorn [1999] for an extensive overview.

We choose to use silhouettes because of the same reasons mentioned in [Agarwal and Triggs, 2004]: they can be extracted relatively robustly from images; they are insensitive to variations in surface such as color and texture; and they encode a great deal of information to recover 3D poses. In addition, they observe that the performance is limited due to artifacts such as shadow and noisy background segmentation; and it is often difficult or impossible to recover certain degrees of freedom due to the lack of depth information. For the encoding of silhouettes, many different shape descriptors are used. In Agarwal and Triggs [2004] it is noted that the use of curve-based shape descriptors, such as Fourier descriptors (FDs) [Zahn and Roskies, 1972] and the Curvature Scale Space (CSS) [Mokhtarian and Mackworth, 1992], is unacceptable since silhouettes can change topology which causes the shape to have discontinuities. We, however, are interested to see if FDs can be used under certain conditions. The advantages of FDs are that they achieve both good representation and good normalization [Zhang and Lu, 2003]. FDs are compact, can be compared with a low-cost Euclidian distance, can be calculated efficiently, are position, rotation and scale invariant and are to some extent insensitive to local noise. We investigate if and under what conditions encoding silhouette boundaries using Fourier descriptors can be used for pose estimation.

3.1 Fourier descriptors

The idea behind Fourier descriptors is to describe a silhouette by a fixed number k of sample points $\{(x_0, y_0), \ldots, (x_{k-1}, y_{k-1})\}$ on the boundary, conform Fig. 1b. The points are usually sampled using equidistant sampling (EDS), where the distance along the silhouette boundary between two consecutive points is constant. These sample points are transformed into complex coordinates $\{z_0, \ldots, z_{k-1}\}$ with $z_i = x_i + y_i\sqrt{-1}$ and are further transformed to the frequency domain using a Discrete Fourier Transform (DFT). The results of this transformation are called the Fourier coefficients, denoted by $\{f_0, \ldots, f_{k-1}\}$.

The coefficients with low index contain information on the general form of the shape and the ones with high index contain information on the finer details of the shape. The first coefficient depends only on the position of the shape and setting it to zero makes the representation position invariant. Rotation invariance is obtained by ignoring the phase information and scale invariance is obtained by dividing the magnitude values of all coefficients by the magnitude of the second coefficient f_1. Since after normalization f_0 is always zero and f_1 is always one, we have $k-2$ unique coefficients given by:

$$\mathbf{FD} = (\frac{|f_2|}{|f_1|}, \ldots, \frac{|f_{k-1}|}{|f_1|})$$

This descriptor is a shape signature and can be used as a basis for similarity and for retrieval. When we sample k points along the boundary, the Fourier descriptor actually has $k-2$ coefficients.

3.2 Sampling points of extreme curvature

Intuitively, points of extreme curvature carry more information. Hands and head can often be found by searching the silhouette boundary for curvature maxima. Similarly, the neck can be found by searching for minima. When points along the boundary are sampled using EDS, it is possible that extreme points are not sampled, especially if a low number of coefficients is used (see for example the feet in Fig. 1b). Therefore we investigate if extreme point sampling (EPS) can be used for matching. A problem when looking for extreme points is the presence of local noise. Therefore we first compute the DFT of all boundary points. Next we set all but the first 200 FDs to zero to eliminate local noise. Now we calculate the number of extreme points m. If we find more than k extreme points, we iteratively lower the number of FDs until we find a number m of extreme points that is not larger than k. In case we have $m < k$ extreme points, we sample $k-m$ additional points along the boundary such that the summed squared distance between all k pairs of subsequent points on the silhouette is minimized. The result of EPS for is shown in Fig. 1c.

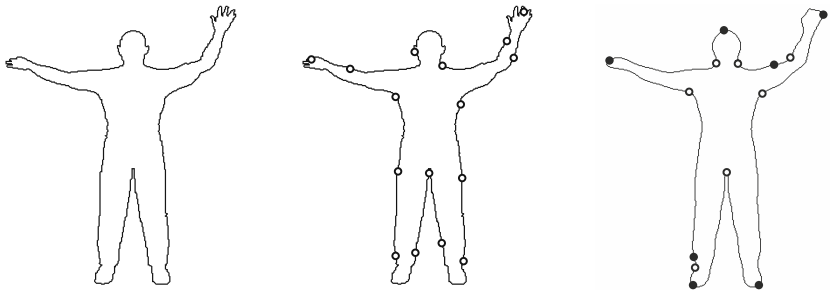
Note that, instead of using point sampling, we could also have taken the entire silhouette boundary and transform all points to the frequency domain. By selecting only the first $k-2$ descriptors, we would have shape signatures of the same dimensionality. However, this approach does not enable us to select

specific boundary points. In informal experiments with this approach, we report error scores very similar to those for EDS. For the sake of clarity, we omit these results from our experiment and discussion.

3.3 Shape similarity based on Fourier descriptors

Now consider two shapes indexed by Fourier descriptors **FD1** and **FD2**. Since both Fourier descriptors are $(k-2)$-dimensional vectors, we can use the Euclidian distance d as a similarity measure between the two shapes:

$$d = \sqrt{(\sum_{j=0}^{k-3} |\mathbf{FD1}_j - \mathbf{FD2}_j|^2)}$$



(a) (b) (c)

Fig. 1: (a) Example silhouette, (b) 16 points sampled along the silhouette boundary with EDS, (c) 16 points sampled using EPS along the boundary of a silhouette reconstructed using 84 Fourier coefficients. Maxima are denoted with a closed circle, minima with an open circle.

4 Pose estimation using Fourier descriptors

We want to recover poses from images. We could approximate the mapping from image space to pose space functionally (see for example Agarwal and Triggs [2004]) but estimating function parameters is difficult given the high non-linearity of the mapping. Instead, we choose to represent the pose space by a finite number of exemplars that sparsely cover the pose domain. For each exemplar, we have a corresponding description in image space, the Fourier descriptor. To recover the pose of a new image, we compare the image description with the descriptions in the database. The estimated pose is the pose that corresponds to the exemplar with the most similar image description. However, this restricts us to recover only poses that are in our example set. Instead, we take the n closest matches and interpolate the poses. In Section 5.3, we determine a suitable value for n.

5 Experiment results and discussion

This section describes the setup and results of the experiments. The example database and test sets are described in Section 5.1 and 5.2 respectively. In Section 5.3, a suitable value for n, the number of poses that are interpolated is determined. The experimental results are presented in Section 5.4 and discussed in Section 5.5.

5.1 Example database

Our example database contains 46,656 silhouette images with their corresponding poses and viewpoint. All combinations of the values given in Tab. 1 are stored in the database. These poses sparsely cover the entire considered pose domain. For simplicity, we excluded the head rotations and assumed a static rotation around the x-axis (*elevation*) of 10°. The rotation around the y-axis is a view parameter and assumes that the presenter addresses the audience (and the camera) with a margin of 80° on both sides. A shoulder twist is a rotation around the upper arm, the front-back rotation is performed in the hand-elbow-shoulder plane and the bend rotation is perpendicular to the other two rotations. Note that our pose domain covers only a small part of all physically possible human poses. This limits the number of exemplars but the approach we take here can also be used for full body pose recovery without loss of generality.

For each pose, a corresponding silhouette is generated using Curious Labs' `POSER` 5. The default `POSER` P5 man was used, with the inverse kinematics settings turned off. We calculated the Fourier descriptors using both EDS and EPS and for descriptor lengths between 8 and 128. Generation of the images and descriptors was performed offline, although it is possible to add examples to the database online.

Rotation	Angle values (°)
Right shoulder twist	{ -90, -45, 0 }
Right shoulder bend	{ -40, 0, 40, 80 }
Right shoulder front-back	{ 0, 45, 90 }
Right forearm bend	{ 0, 40 }
Left shoulder twist	{ -90, -45, 0 }
Left shoulder bend	{ -80, -40, 0, 40 }
Left shoulder front-back	{ -90, -45, 0 }
Left forearm bend	{ -40, 0 }
Rotation around y-axis	{ -80, -60, -40, -20, 0, 20, 40, 60, 80 }

Tab. 1: Degrees of freedom with the angle value ranges that are stored in the example image database.

5.2 Test sets

We wanted to measure the pose recovery performance for different kinds of silhouette shape deformations. Therefore we generated four test sets (**T1** ... **T4**), each with a different deformation. Each test set contains 1,000 images and $\mathbf{T}_{i,j}$, the j^{th} image in test set $\mathbf{T}_i$ $(1 \leq i \leq 4)$ corresponds to the same pose $p_j \in \mathbb{R}^9$. Each degree of freedom in pose p_j is chosen within the ranges given in Tab. 1.

- **T1** contains the `POSER` P5 default man.
- **T2** contains the `POSER` P5 default woman, who has different body dimensions.
- **T3** contains the `POSER` P5 default man but viewed from a different angle. The elevation is 20° instead of 10°. This rotation is not part of the pose description and the test set serves to see if small changes in viewpoint can be handled correctly.
- **T4** contains the `POSER` P5 default man but with noise generated on the boundary. This set allows us to see how much effect noise has on the recovery performance. Noise was added to a sample point by adding a vector of length l in the direction of the contour normal in the sampled point. l is sampled from a normal distribution with zero-mean and a variance of 2% of the silhouette height.

Example silhouettes are shown in Fig. 3. There are no images generated for **T4**, the contours are obtained by adding noise to points sampled along the boundary of images from **T1**.

5.3 Determining the value for n

The average sum of distances over all 9 DOF between a pose with values for each DOF randomly chosen from an uniform distribution over the ranges given in Tab. 1 and the closest pose in the example database is 180°. By considering not only the closest exemplar but the n closest and interpolating the corresponding poses, we could lower this error. We use a normalized weighted n-best interpolation. In this section, we determine the value for n, the number of exemplars that are used for interpolation. We use the example database and **T1** with 64 points sampled with both EDS and EPS to empirically determine a suitable value for n. The results are shown in Fig. 2. The plotted values are the average sum of estimation errors in ° for all DOF over all images in **T1**.

It is clear that by choosing n small, the average error is high. For $n \geq 25$, the error is more or less constant. Therefore, for the experiments described in this paper, we use a normalized weighted n-best interpolation with the $n = 25$ best matches.

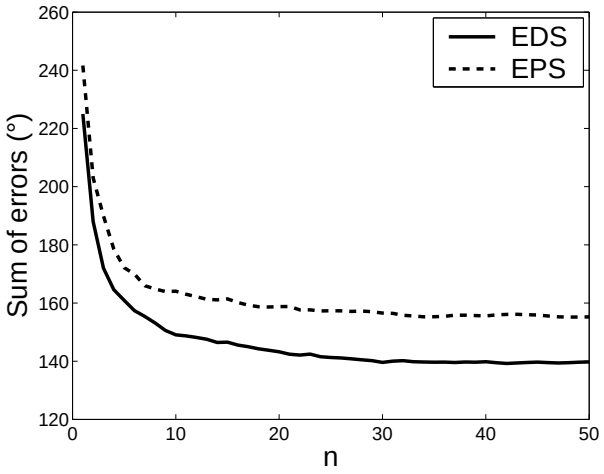


Fig. 2: Average sum of estimation errors in ° for all DOF over **T1** using an n-best weighted average, sampled using EDS and EPS with $k = 64$ sampled points.

5.4 Experiment results

We performed experiments with the four test sets, sampled using EDS and EPS and with a number of sampled points between 8 and 128. The results for EDS are summarized in Tab. 2, for EPS in Tab. 3. These tables show the sum of errors over all 9 DOF, averaged over the entire test set. Note that the example database and test sets are always sampled with the same method. Fig. 4 and 5 show silhouettes that are reconstructed using `POSER` from the estimates of EDS and EPS respectively, and $k = 64$.

Test set	Mean error angle distance (in °)				
	$k = 8$	$k = 16$	$k = 32$	$k = 64$	$k = 128$
T1	150.31	144.88	145.27	144.77	145.83
T2	164.69	164.33	164.90	164.86	165.45
T3	155.33	151.40	150.99	149.87	150.44
T4	150.19	145.89	145.83	145.33	145.16

Tab. 2: Mean estimation error in ° for test sets **T1** . . . **T4**, sampled using EDS with different descriptor lengths.

5.5 Discussion

The baseline for the sum of errors of all DOF for a single image is 280°. This is the sum over all DOF of expected values for the distance between two numbers

Test set	Mean error angle distance (in °)				
	$k = 8$	$k = 16$	$k = 32$	$k = 64$	$k = 128$
T1	156.89	154.87	156.52	149.62	149.96
T2	172.17	173.46	172.78	165.96	166.00
T3	158.97	163.13	159.68	154.42	154.90
T4	156.88	155.82	156.87	149.67	150.18

Tab. 3: Mean estimation error in ° for test sets **T1** . . . **T4**, sampled using EPS with different descriptor lengths.

randomly selected from a uniform distribution between the ranges for a single DOF. It is clear that all summed errors are significantly lower than this baseline.

Our first observation is that the differences in performance between EDS and EPS are small. The error scores of EDS are slightly lower than those of EPS, especially when less than 64 Fourier coefficients are used. Apparently, for EPS the smoothing that is applied by describing the contour with less coefficients in order to obtain the extreme points we discard useful information. As k increases, m will also increase and more detail of the contour is preserved. This leads, ultimately, to an increased matching performance.

Decreasing the number of coefficients k has only a small effect on the estimation error. For both EDS and EPS, we see a decrease in error between 8 and 128 coefficients of approximately 5°. We can describe our silhouettes with as few as 8 coefficients and obtain average summed errors that are comparable with those obtained by Shakhnarovich et al. [2003].

If we look at the test sets, we also notice small differences. The results of **T1** and **T4** are comparable, which indicates that our approach is robust to small variations in silhouette shape due to noise. **T3** scores only approximately 5° worse than **T1**, which suggests that variations in viewpoint can be handled to some extent. We notice that **T2** scores about 10% worse than **T1**. The variations in body dimensions tend to result in inaccurate estimations.

Tab. 4 and Tab. 5 show the average error per joint over the entire test set using EDS and EPS, respectively. We see relatively low errors for the forearm bend. This could also be due to the limited range of only 40°, which would yield a mean estimation error of 13.3° for random guesses. The error for the shoulder twist is a little higher than the other shoulder rotations because it is very difficult to estimate this rotation when the elbow is completely stretched. We report a low error for the rotation around the y-axis but also note that we have example images every 20° for this DOF instead of every 40 or 45°. We performed an additional experiment where we removed all poses that had a value for the rotation around the y-axis of -60, -20, 20 and 60°. This is a reduction of 44% of the pose space. The results are summarized in Tab. 6. Although other joints also have slightly higher average errors, we report a substantial increase in the error for the rotation around the y-axis. This leads us to believe that adding more exemplars in the pose domain will lower the error scores.

6 Conclusion and future work

We investigated whether Fourier descriptors can be used to encode a silhouette boundary for pose estimation. Descriptors with a number low number of coefficients were calculated using equidistant sampling (EDS) and extreme point sampling (EPS). An example image set with 46,656 silhouette images of a man in various poses was used to recover a 9 DOF pose. We performed tests with deformed shapes to test the robustness against variations in body dimensions, viewpoint and noise. Body poses were recovered by interpolating the poses corresponding to the n-best matches. The average error per joint was approximately 16-17° for EDS and slightly higher for EPS. Increasing the number of descriptors did not have any influence on the performance. Adding noise to the silhouette boundary did not affect the recovery performance and small changes in viewpoint increased the error only slightly. However, we obtained higher error scores when recovering poses using silhouettes of a woman, who has different body dimensions. Adding more exemplars in the pose domain is likely to result in a lower estimation error, for all deformations.

Future work will aim at finding a robust descriptor that is able to cope with variations in body dimensions and large-scale occlusions. This allows the work to be used for pose recovery in realistic situations. We also plan to incorporate dynamics to improve our estimation results.

References

Ankur Agarwal and Bill Triggs. 3D human pose from silhouettes by relevance vector regression. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, volume 2, pages 882–888, Washington, DC, June 2004.

Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, April 2002.

Matthew Brand. Shadow puppetry. In *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV'99)*, volume 2, pages 1237–1244, Kerkyra, Greece, September 1999.

Quentin Delamarre and Olivier Faugeras. 3D articulated models and multi-view tracking with silhouettes. In *Proceedings of the 7th IEEE International Conference on Computer Vision (ICCV'99)*, volume 2, pages 716–721, Kerkyra, Greece, September 1999.

Jonathan Deutscher, Andrew Blake, and Ian Reid. Articulated body motion capture by annealed particle filtering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'00)*, volume 2, pages 126–133, Hilton Head Island, SC, June 2000.

Ahmed M. Elgammal and Chan-Su Lee. Inferring 3D body pose from silhouettes using activity manifold learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, volume 2, pages 681–688, Washington, DC, June 2004.

Dariu M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding (CVIU)*, 73(1):82–92, January 1999.

Nicholas R. Howe. Silhouette lookup for automatic pose tracking. In *Proceedings of the IEEE Workshop on Articulated and Nonrigid Motion*, volume 1, pages 15–22, Los Alamitos, CA, June 2004.

Ming-Kuei Hu. Visual pattern recognition by moment invariants. *IRE Transactions Information Theory*, 8(2):179–187, February 1962.

Thomas B. Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding (CVIU)*, 81(3):231–268, March 2001.

Farzin Mokhtarian and Alan K. Mackworth. A theory of multi-scale, curvature-based shape representation for planar curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8):789–805, August 1992.

Greg Mori and Jitendra Malik. Estimating human body configurations using shape context matching. In *Proceedings of the European Conference on Computer Vision (ECCV'02)*, number 2352 in Lecture Notes in Computer Science, pages 666–680, Copenhagen, Denmark, May 2002.

Liu Ren, Gregory Shakhnarovich, Jessica K. Hodgins, Hanspeter Pfister, and Paul Viola. Learning silhouette features for control of human motion. In *Proceedings of the SIGGRAPH 2004 Conference on Sketches & Applications*, Los Angeles, CA, August 2004.

Rómer E. Rosales and Stan Sclaroff. Inferring body pose without tracking body parts. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'00)*, volume 2, pages 721–727, Hilton Head Island, SC, June 2000.

Rómer E. Rosales, Matheen Siddiqui, Jonathan Alon, and Stan Sclaroff. Estimating 3D body pose using uncalibrated cameras. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, volume 1, pages 821–827, Kauai, HI, December 2001.

Gregory Shakhnarovich, Paul Viola, and Trevor Darrell. Fast pose estimation with parameter-sensitive hashing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'03)*, volume 2, pages 750–759, Nice, France, October 2003.

Hedvig Sidenbladh, Michael J. Black, and David J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *Proceedings of the European Conference on Computer Vision (ECCV'00)*, volume 2492 of *Lecture Notes in Computer Science*, pages 702–718, Dublin, Ireland, June 2000.

Christian Sminchisescu and Bill Triggs. Kinematic jump processes for monocular 3D human tracking. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)*, volume 1, pages 69–76, Madison, WI, June 2003. IEEE Computer Society.

Josephine Sullivan and Stefan Carlsson. Recognizing and tracking human action. In *Proceedings of the European Conference on Computer Vision (ECCV'02) - volume 1*, number 2350 in Lecture Notes in Computer Science, pages 629–644, Copenhagen, Denmark, May 2002.

Remco C. Veltkamp and Michiel Hagedoorn. State-of-the-art in shape matching. Technical Report UU-CS-1999-27, Utrecht University, Utrecht, The Netherlands, September 1999.

Charles T. Zahn and Ralph Z. Roskies. Fourier descriptors for plane closed curves. *IEEE Transactions on Computers*, C-21(3):269–281, March 1972.

Dengsheng Zhang and Guojun Lu. A comparative study of curvature scale space and fourier descriptors for shape-based image retrieval. *Journal of Visual Communication and Image Representation*, 14(1):39–57, March 2003.

Rotation	Mean (SD) error angle distance in °			
	T1	**T2**	**T3**	**T4**
Right shoulder twist	19.91 (14.07)	21.20 (14.84)	20.21 (14.75)	19.79 (13.91)
Right shoulder bend	21.44 (17.96)	25.06 (19.52)	22.55 (17.97)	21.50 (17.81)
Right shoulder front-back	16.01 (13.13)	18.00 (14.16)	16.59 (13.29)	16.20 (13.08)
Left shoulder twist	19.76 (13.96)	20.80 (14.96)	18.97 (14.33)	19.79 (13.77)
Left shoulder bend	22.00 (18.31)	25.77 (18.60)	22.58 (17.59)	22.17 (18.20)
Left shoulder front-back	16.20 (12.97)	17.99 (14.26)	17.13 (13.36)	16.38 (13.04)
Right forearm bend	10.72 (7.87)	11.31 (8.23)	11.01 (7.98)	10.66 (7.88)
Left forearm bend	10.89 (7.52)	11.37 (7.85)	11.48 (8.03)	10.85 (7.54)
Rotation around y-axis	7.85 (8.04)	13.37 (11.77)	9.36 (8.74)	7.98 (8.19)
Total	144.77 (46.75)	164.86 (51.21)	149.87 (44.08)	145.33 (46.71)
Average	16.09	18.32	16.65	16.15

Tab. 4: Mean and standard deviation of the estimation errors per DOF for in ° for test sets **T1** . . . **T4**, sampled using EDS and $k = 64$ coefficients.

Rotation	Mean (SD) error angle distance in °			
	T1	**T2**	**T3**	**T4**
Right shoulder twist	19.91 (14.12)	21.65 (15.13)	20.97 (15.70)	19.77 (13.84)
Right shoulder bend	22.70 (18.30)	25.21 (19.84)	23.67 (18.34)	22.72 (18.21)
Right shoulder front-back	17.01 (13.47)	18.74 (14.73)	17.57 (14.09)	16.97 (13.31)
Left shoulder twist	20.09 (14.74)	20.68 (14.49)	19.49 (14.83)	20.27 (14.80)
Left shoulder bend	23.01 (18.97)	25.17 (18.36)	23.21 (18.30)	23.21 (18.83)
Left shoulder front-back	17.22 (13.71)	17.67 (13.49)	17.50 (13.67)	17.15 (13.65)
Right forearm bend	10.72 (7.55)	11.38 (8.00)	11.09 (7.77)	10.63 (7.49)
Left forearm bend	10.62 (7.40)	11.41 (7.65)	11.26 (7.56)	10.57 (7.38)
Rotation around y-axis	8.34 (9.03)	14.07 (12.11)	9.66 (9.07)	8.39 (9.19)
Total	149.62 (48.08)	165.96 (49.84)	154.42 (45.76)	149.67 (47.75)
Average	16.62	18.44	17.16	16.63

Tab. 5: Mean and standard deviation of the estimation errors per DOF for in ° for test sets **T1** . . . **T4**, sampled using EPS and $k = 64$ coefficients.

Rotation	Mean (SD) error angle distance in °			
	T1	**T2**	**T3**	**T4**
Right shoulder twist	20.93 (14.67)	22.70 (15.18)	21.05 (15.35)	20.89 (14.42)
Right shoulder bend	23.20 (19.26)	25.77 (20.63)	23.77 (18.41)	23.17 (19.39)
Right shoulder front-back	16.35 (13.38)	18.72 (13.92)	17.02 (13.22)	16.32 (13.24)
Left shoulder twist	20.10 (14.85)	21.16 (15.33)	19.13 (14.36)	20.09 (14.50)
Left shoulder bend	22.35 (18.40)	26.58 (19.43)	22.93 (17.74)	22.17 (17.99)
Left shoulder front-back	16.75 (13.62)	18.76 (14.64)	17.44 (13.38)	16.85 (13.66)
Right forearm bend	11.19 (7.75)	11.15 (8.18)	11.00 (7.70)	11.16 (7.72)
Left forearm bend	11.16 (7.64)	11.87 (8.16)	11.65 (8.13)	10.99 (7.61)
Rotation around y-axis	11.24 (10.11)	17.78 (14.45)	12.27 (10.58)	11.25 (9.85)
Total	153.27 (49.73)	174.51 (53.41)	156.27 (44.82)	152.88 (48.62)
Average	17.03	19.39	17.36	16.99

Tab. 6: Mean and standard deviation of the estimation errors per DOF for in ° for test sets **T1** . . . **T4**, sampled using EDS and $k = 64$ coefficients. The number of exemplars is reduced by 44%.

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

Fig. 3: The rows contain images from **T1**, **T2** and **T3**, respectively. (a-d) correspond to pose p_j, $j \in \{1 \dots 4\}$

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

Fig. 4: The rows contain images reconstructed from the estimates from **T1**...**T4**, using EDS with $k = 64$ points. (a-d) correspond to pose p_j, $j \in \{1 \dots 4\}$

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

Fig. 5: The rows contain images reconstructed from the estimates from **T1**...**T4**, using EPS with $k = 64$ points. (a-d) correspond to pose p_j, $j \in \{1 \dots 4\}$