Comparison of Silhouette Shape Descriptors for Example-based Human Pose Recovery

Ronald Poppe and Mannes Poel [*]
University of Twente, Dept. of Computer Science, Human Media Interaction Group
P.O. Box 217, 7500 AE Enschede, the Netherlands

E-mail: {poppe,mpoel}@ewi.utwente.nl

Abstract

Automatically recovering 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 three shape descriptors that are used in the encoding of silhouettes: Fourier descriptors, shape contexts and Hu moments. An example-based approach is taken to recover upper body poses from these descriptors. We perform experiments with deformed silhouettes to test each descriptor's robustness against variations in body dimensions, viewpoint and noise. It is shown that Fourier descriptors and shape context histograms outperform Hu moments for all deformations.

1 Introduction

Being able to estimate human poses from visual input automatically 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 multi-valued 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 pose recovery 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 is often difficult. *Learning-based* 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 or 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 [1]. 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 three shape descriptors that are used in the encoding of silhouettes: Fourier descriptors, shape contexts and Hu moments. An example-based approach is taken to recover upper body poses from each of these three 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 three shape 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. 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. In Section 3 the three shape descriptors that

[*] This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-155), and is part of the ICIS program. ICIS is sponsored by the Dutch government under contract BSIK03024.

are compared in this paper are discussed in more detail, 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 [5]. Current research can roughly be divided into *model-based* and *learning-based* approaches. Model-based approaches [3, 12] 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 [13]. Learning-based approaches [1, 11] 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 used throughout literature. Silhouettes and edges are used the most, because they can be easily extracted and are, to some extent, lighting invariant.

Howe [6] uses silhouettes which are matched to a collection of known poses using turning angle and Chamfer distance. Agarwal and Triggs [1] encode the silhouette boundary using shape contexts and use Bayesian non-linear regression to recover a 54 DOF body pose with high accuracy. Elgammal and Lee [4] learn view-based activity manifolds from silhouettes. Mappings from silhouette to activity manifold and from activity manifold to pose are learned from training data.

Edges contain more information but are also more sensitive to texture. Mori and Malik [9] 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. Shakhnarovich *et al.* [11] 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.* [3] use a particle filter to propagate movements in time. Another approach is to learn the dynamics of human movement from training samples [1, 4, 12]. Although this often leads to more stable and accurate estimation results, it also puts a strong prior on the movements that can be recovered.

Our work is related to these approaches in that we recover poses from visual input. However, 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 descriptors, see [14]. We selected Fourier descriptors, shape contexts and Hu moments because of their different characteristics and their use in previous work.

3 Silhouette shape descriptors

An ideal descriptor for our pose recovery problem would be able to distinguish between different body poses while being able to generalize over body dimensions, variations in viewpoint and local boundary noise. Here we focus on three descriptors: Fourier descriptors, shape contexts and Hu moments. Each of the descriptors has distinct characteristics, which are described in the next sections.

3.1 Fourier descriptors

The idea behind Fourier descriptors [15] is to describe a silhouette by a fixed number n of sample points $\{(x_0, y_0), \ldots, (x_{n-1}, y_{n-1})\}$ on the boundary. Since Fourier descriptors can describe only a single closed-curve shape, we only sample along the external boundary. Sampling can be done randomly but in practice, equidistant sampling is preferred to make sure the sampling is more uniform. Poppe and Poel [10] experimented with sampling extreme boundary points but found no significant difference in performance. The sample points are transformed into complex coordinates $\{z_0, \ldots, z_{n-1}\}$ 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 $\{f_0, \ldots, f_{n-1}\}$.

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 $n-2$ unique coefficients given by:

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

This descriptor is a shape signature and can be used as a basis for similarity and for retrieval. When we deal with Fourier descriptors of length n, we actually use only $n-2$ coefficients. It is clear that Fourier descriptors describe a shape globally. The finer details are filtered out.

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

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

3.2 Shape contexts

Shape contexts [2] are rich local descriptors of sampled internal or external boundary points. A shape context describes the spatial locations of the other $n-1$ sampled points in a histogram. Each histogram bin is uniform in log-polar space. The shape context is parameterized by the number of radial bins ϕ, number of log-distance bins r, and inner and outer distance boundary r_{inner} and r_{outer}. A single shape is described by the shape contexts of all n sampled points.

Shape contexts are translation invariant by definition since the distances are measured with respect to the sampled point. Scale invariance is obtained by normalizing the distances between points with the mean distances between all n^2 point pairs. Rotation invariance for each point can be achieved by taking as a reference frame the tangent vector of the point, instead of the positive x-axis.

We use the default values $\phi = 12$ and $r = 5$ for the number of bins. Agarwal and Triggs [1] lower r_{outer} and suggest that by normalizing the shape context, robustness against shape deformations is increased. Indeed, in informal experiments we notice a small increase in performance over the default r_{outer} and without normalization. In our experiments, we set r_{outer} to the mean distance between all points. The number of shape context bins over an entire image is $r\phi n$. This number gets large for a high number of sampled points, which is inconvenient for storage and comparison. Therefore we cluster the shape context space into m clusters. We will use $m = 100$ clusters throughout the paper. The contribution η_i of the i^{th} ($1 \leq i \leq m$) cluster $\mathbf{C}_i$ to shape context **SC** is calculated using soft voting:

$$\eta_i(SC) = \frac{min_{r=1\ldots m}|\mathbf{SC} - \mathbf{C}_r|^2}{|\mathbf{SC} - \mathbf{C}_i|^2}$$

The η_i's are summed over the shape contexts of all n sampled points of a shape and normalized to unit length. This allows us to compare two shapes with a different number of sampled points with the Euclidian distance between two m-dimensional vectors, but we lose all spatial information [8]. This histogram of shape context center contributions forms a descriptor that captures both the global and local characteristics of the shape. Agarwal and Triggs [1] also use vector quantization to vote softly into the centers but instead use Gaussian weights. Since we have normalized our shape contexts, all points lie on a 60D hypersphere, which makes the covariance matrix C nearly singular. We could solve this problem by using $C^{'} = C + \lambda I$ instead but choosing the value of λ is non-trivial.

3.3 Hu moments

In contrast to Fourier descriptors and shape contexts, moments are region-based descriptors. Hu [7] derived 7 orthogonal invariant moment descriptors of order 2 and 3. The first 6 descriptors encode a shape with invariance to translation, scale and rotation. The 7^{th} descriptor ensures skew invariance, which enables us to distinguish between mirrored images. A Hu moment **HU** is denoted by its descriptors $\mathbf{HU}_i$ ($1 \leq i \leq 7$).

The descriptors are of different orders and therefore the difference between two Hu moments cannot be computed using Euclidian distance. Instead, we calculate the difference d^2 between two Hu moments **HU1** and **HU2** as:

$$d^2 = |\mathbf{HU1} - \mathbf{HU2}|\Sigma^{-1}|\mathbf{HU1} - \mathbf{HU2}|^T$$

In this equation, $\mathbf{\Sigma}$ is the covariance matrix that is calculated over the entire example set.

4 Pose recovery

We want to recover poses from images. We could approximate the mapping from image space to pose space functionally (see for example [1]) 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. 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. Instead of taking the single best match, one could also take the n-best matches and interpolate the poses. Similar to [10], we use a normalized weighted n-best interpolation of the 25 best matches.

5 Experimental results & 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. The experimental results are presented in Section 5.3 and discussed in Section 5.4.

5.1 Example database

Using Curious Labs' POSER we constructed a database with 46,656 silhouette images of the POSER P5 default man in various poses and viewed from different angles. The 9 degrees of freedom and their possible values are summarized in Table 1.

Table 1. Degrees of freedom with the angle values (in °) that are stored in the example database

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

We only consider joint rotations in the upper body to limit the number of needed exemplars somewhat, but the approach could be used for full body poses without loss of generality. Furthermore, we excluded the head rotations and assumed a static rotation around the x-axis (*elevation*) of 10°. Shoulder twist is the 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.

We calculated the image descriptors (Fourier descriptors, histograms of shape contexts and Hu descriptors) offline. These descriptors, together with their corresponding pose descriptions form the example database.

5.2 Test sets

We used four different test sets ($T1 \ldots T4$) to measure the pose recovery performance for different kinds of shape deformations. Each test set contains 1,000 poses, and the j^{th} image each the test set $T_{i,j}$ ($1 \leq i \leq 4$) 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 Table 1. Two example poses, rendered in each test set, are shown in Figure 1.

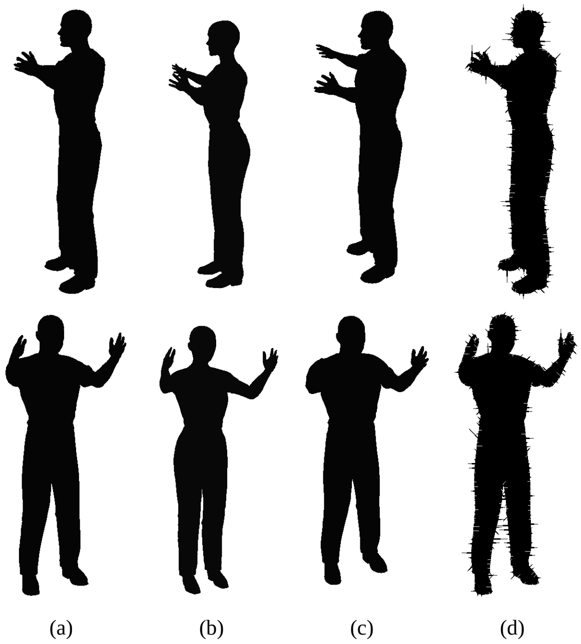

Figure 1. (a-c) Example pose rendered in T1, T2 and T3 respectively. (d) The filled polygon of contour points with noise added. Images of this type are only used in T4 for the experiments with Hu moments. The rows each represent a randomly chosen sample.

- **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. For the Hu moment test set, we created images where we filled the polygon of all points with added noise (Figure 1d).

5.3 Experiment results

We estimated the pose from each of the three shape descriptors individually. The means and standard deviations of the sum of the 9 joint rotation errors of the experiments are shown in Table 2, 3 and 4 for Fourier descriptors, shape context histograms and Hu moments respectively. Note that we carried out the Fourier descriptor and shape context experiments with three different values for the number of sampled points n.

5.4 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 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.

Table 2. Mean and standard deviation (in °) of the sum of estimation errors for the Fourier descriptor experiments

Test	FD16	FD64	FD256
	Mean (SD)	Mean (SD)	Mean (SD)
T1	154.59 (46.96)	153.24 (50.41)	153.53 (50.59)
T2	164.29 (49.47)	163.33 (52.83)	163.08 (52.84)
T3	152.68 (43.39)	150.51 (44.21)	150.98 (44.27)
T4	157.47 (46.93)	154.76 (51.44)	153.88 (50.20)

Table 3. Mean and standard deviation (in °) of the sum of estimation errors for the shape context histogram experiments

Test	FD16	FD64	FD256
	Mean (SD)	Mean (SD)	Mean (SD)
T1	163.42 (44.27)	147.79 (47.99)	150.45 (52.43)
T2	174.07 (47.42)	163.39 (53.40)	164.56 (55.97)
T3	169.36 (45.62)	152.33 (45.62)	153.92 (46.09)
T4	169.51 (45.69)	150.09 (47.38)	153.12 (52.54)

Table 4. Mean and standard deviation (in °) of the sum of estimation errors for the Hu moment experiments

Test	Mean (SD)
T1	183.24 (55.90)
T2	194.80 (58.95)
T3	199.17 (50.28)
T4	195.40 (58.54)

The first observation is that the differences between shape descriptors are small. The Hu moments score a little lower than the contour-based descriptors. The errors for the Fourier descriptor experiment are in line with the results reported in [10]. Two shape contexts are usually matched using bipartite graph matching, where an additional penalty term is introduced for alignment error between the two contexts. Since we describe a silhouette as a single histogram, this spatial arrangement information is lost. This might explain the fact that the results for the rich shape contexts are comparable to those of the Fourier descriptors.

We also note small differences between the test sets of a single descriptor. Compared to **T1**, both contour-based descriptors show good results on **T3** and **T4**. We may conclude that both descriptors are robust against small variations in shape due to viewpoint changes and noise. We report slightly higher error scores in the **T2** set for the contour-based descriptors. Remarkably enough, this effect is not present in the Hu moment experiments.

The number of sampled points n in the experiments with the contour-based descriptors does not make a big difference. This was already observed for Fourier descriptors [10]. For the histograms of shape contexts, this is due to the fact that we used soft voting. In an additional experiment where we used hard clustering instead, we reported a sum of errors that was approximately 10° higher for $n = 16$. The results for $n = 64$ and $n = 256$ are almost similar to those obtained using soft voting.

Table 5 shows the mean estimation error on **T1**. For the contour-based descriptors, n was set to 64. 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 a random guess. 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 forearm is completely stretched. This situation occurs in 50% of all samples. Furthermore, 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°. The results of three additional experiments are summarized at the bottom of Table 5. We see that voting hard into shape context centers performs equally well as soft voting. The experiments with the default shape context parameters that have been used in [2] for shape matching result in slightly higher errors than the experiments with an our adjusted r_{outer}. Not surprisingly, ignoring Hu moments' skew invariance results in a dramatically higher error for the rotation around the y-axis.

Table 5. Average joint estimation errors (in °) for Fourier descriptors, shape context histograms and Hu moments with different sets of parameters

Test conditions	Right shoulder			Left shoulder			Forearm bend		Around
	Twist	Bend	Front-back	Twist	Bend	Front-back	Right	Left	y-axis
Shape context soft vote	20.61	22.79	16.09	20.44	22.59	16.20	10.97	10.84	7.27
Fourier descriptor	21.03	23.21	17.34	20.40	23.88	17.04	11.17	10.42	8.76
Hu moment	24.13	29.94	20.64	23.36	29.29	20.79	11.30	11.74	12.05
Shape context hard vote	21.15	22.86	16.43	20.59	22.66	16.65	10.65	10.42	6.70
Shape context default params	20.75	25.82	18.10	20.43	25.98	19.29	10.04	10.48	13.11
Hu moment skew variant	23.98	30.92	22.02	23.43	30.66	21.40	11.13	11.20	40.11

6 Conclusions & future work

We compared silhouette shape descriptors for vision-based pose recovery. An example set of 46,656 images, each representing a different pose, was encoded using Fourier descriptors, shape context histograms and Hu moments. These were stored, together with their corresponding poses. We performed tests with deformed shapes to test each descriptor's robustness against variations in body dimensions, viewpoint and noise. It is shown that Fourier descriptors and shape context histogram outperform Hu moments for all deformations.

Future work will aim at finding a robust descriptor that is also able to cope with large-scale occlusions. This allows the work to be used for pose recovery in realistic situations.

References

[1] A. Agarwal and B. 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.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, April 2002.

[3] J. Deutscher, A. Blake, and I. 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.

[4] A. M. Elgammal and C.-S. 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.

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

[6] N. 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.

[7] M.-K. Hu. Visual pattern recognition by moment invariants. *IRE Transactions Information Theory*, 8(2):179–187, February 1962.

[8] G. Mori, S. Belongie, and J. Malik. Efficient shape matching using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11):1832–1837, November 2005.

[9] G. Mori and J. 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.

[10] R. Poppe and M. Poel. Example-based pose estimation in monocular images using compact fourier descriptors. Technical Report TR-CTIT-05-49, University of Twente, Enschede, The Netherlands, October 2005.

[11] G. Shakhnarovich, P. Viola, and T. 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.

[12] H. Sidenbladh, M. J. Black, and D. 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.

[13] C. Sminchisescu and B. 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.

[14] R. C. Veltkamp and M. Hagedoorn. State-of-the-art in shape matching. Technical Report UU-CS-1999-27, Utrecht University, Utrecht, The Netherlands, September 1999.

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