

STATISTICAL REGRESSION METHOD OF SHAPE ANALYSIS, WITH APPLICATION TO CLASSIFICATION OF CROSS-SECTIONS OF CARBON FIBERS

ALEŠ LINKA, PETR VOLF

Abstract. The contours of textile fibers cross-sections have as a rule the shape of deformed circle. Unfolded contours can then be described as a smooth periodical curve contaminated with local nonregularities, and can therefore be analyzed with the aid of the statistical regression model. The present paper applies such an approach to the comparison of shapes of cross-sections for several types of carbon fibers. The corresponding regression curve is constructed as the combination of trigonometric functions, its complexity is optimized by methods of mathematical statistics. It is also shown, by random sampling, that the parameters of the model correspond to different deformations of circular contour. On this basis, the method is proposed for the discrimination between the heat-treated and untreated fibers.

Резюмэ. В статье предлагается метод статистического анализа формы 2D объэктов. Метод пользуется функциональной моделью контура объэкта посьле его развертки. Деформациа контура описана параметрами модели. В качестве примера решается задача характеризации и классификации текстилных волокон на основе контуров их сечений.

1. Introduction, mathematical analysis of shapes

The mathematical methods of analysis of shapes have, in recent decade, attracted rather wide attention. The remarkable development has been achieved particularly in the field of image processing, of stereology, and also in the area of stochastic models and application of statistical data analysis. In the present paper we shall deal with an application of statistical methods to the description and classification of shapes of cross-sections of carbon fibers. The objective is to find the most informative features of these shapes and to describe the differences between several types of fibers, namely between the heat-treated and untreated ones.

The temperature exposition of fibers during technological processing of simple or composite materials produces very often changes both in size and shape of their cross-sections as a consequence of reorganization of their internal structure. Well known is for instance the irreversible shrinkage of textile fibers by heating. This phenomenon is observed also at high performance fibers used as composite reinforcement, for example at carbon fibers, which are exposed to graphization temperature at about 2900 $°C$. In the case of carbon fibers the identification of these small changes may be very difficult and the mathematical methods are the useful tool of such an analysis.

2000 *Mathematics Subject Classification.* 62H35 62P30.

Key words and phrases. Carbon fibers, cross-sections, analysis of shapes, statistical methods, regression model.

The authors express their appreciation to Mrs. Bohumila Košková for providing the experimental data, which were obtained within Grant project No. 106/99/0096 of Grant Agency of Czech Republic. The research has been supported by the project of GA ČR No. 106/01/0565.

Naturally, one of the main characteristics is the size of the cross-section. It can be measured by the area, minimal or maximal distance of object's points, mean diameter, etc. Another set of characteristics describes typical features of the shape. One of the most popular approaches suitable for the problem solved here is based on the concept of deformable template model (Grenander, 1993). It assumes that the object is a result of a (stochastic) deformation of a template and the aim is to describe the deformation mechanism.

In the present paper, we shall use the statistical model describing the unfolded contour of examined object (i.e. the cross-section of a fiber) via the regression model $r(u) = g(u) + \varepsilon$, where u is the angle (from 0 to 2π), $r(u)$ is the length of radius from an appropriately chosen point c (a "center" of the object) to the edge, ε is the random noise and g is the model function. Its form, resp. its parameters, are expected to carry an important information on the shape of deformed object.

In the case considered here, such a 'regression-like' description of the contours of cross-sections is quite adequate because the shapes considered are actually flattened circles (with additional more or less considerable nonregularities). That is why the contours of fiber cross-sections can be unfolded to a curve – function. Such shapes are sometimes called 'star-shaped'. It means that there exists at least one point c inside the object such that the line segments connecting c with each point on the contour are inside the object (see, for instance, Hobolth et al, 2000). Moreover, it will be assumed that the point c can be chosen as the center of gravity of the considered objects – cross-sections.

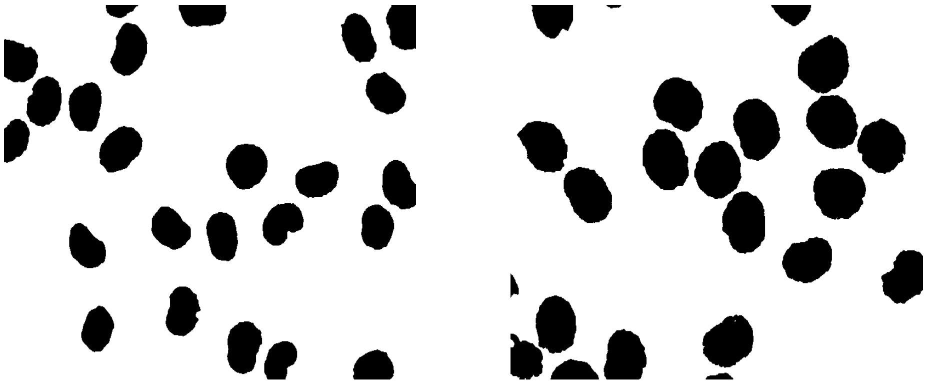
From what has been said it follows that we shall deal with planar objects, and that the location or rotation of the object is not relevant to the purpose of our study, at least in the present stage. For instance, when examining the fibers in a bundle, one should consider the location, too, because the conditions in the center of the bundle may differ significantly from the conditions at points close to its border.

In the follow-up, we shall first present the data, the cross-sections of carbon fibers of two types, and we shall compare their magnitude. Then, the regression model of unfolded contours of cross-sections will be formulated and analyzed. Finally, the simulations will confirm the correspondence between the deformation of shapes and the parameters of the model. We shall also mention the shape analysis method based on the comparison of significant points selected on the object contour – so called landmarks – by means of Procrust analysis.

2. The data, comparison of the magnitude of cross-sections

Two samples of cross-sections of carbon fibers were analyzed. Their images are on Figure 1, they were obtained with the aid of a confocal microscope and CCD camera. The cross-sections of the first type (a) are from the annealed (heat-treated) fibers, the second type (b) is not annealed. We analyzed $N_1 = 16$ items of the first type and $N_2 = 14$ of the second.

It is seen that the shapes have an approximately elliptical form, so that there is no problem to establish their centers of gravity. Further, it is then possible to measure the lengths of radii and to compute the average radius. For the purpose of the analysis the contour of each cross-section was stored as a set of values $\{(x_i, y_i),\ i = 1, \ldots, n_j\}$ in a local coordinate system x, y, where j is the number of cross-section and $n_j \sim 2 \cdot 10^2$ points. The units used for x_i, y_i were the numbers of pixels in the image (one unit = one pixel, in both x and y directions), while the magnitude (diameter) of the real cross-sections was about $10\mu m$. The gravity center $c_j = (c_{xj}, c_{yj})$ of j-th object is given by $c_{xj} = \sum x_i/n_j$, $c_{yj} = \sum y_i/n_j$, coordinates w.r. to this new center are $x'_i = x_i - c_{xj}$, $y'_i = y_i - c_{yj}$, the length of the radius connecting the center (now $(0,0)$) with the contour point (x'_i, y'_i) is $r_i = \sqrt{x'^2_i - y'^2_i}$ and the corresponding angle u_i between the radius-vector and axis x covers the interval $(0, 2\pi)$. More precisely, $u_i = \operatorname{arc\,tg}(y'_i/x'_i)$, shifted by $+\pi$ if $x_i < 0$ and shifted by $+2\pi$ for $x_i \geq 0$, $y < 0$.



(a) (b)

Fig. 1 *The data – cross-sections of heat-treated (a) and untreated (b) carbon fibers.*

Further, denote $R_j = \sum_{i=1}^{n_j} r_i/n_j$ the averaged radius of object j. In such a way we obtained two sets of values $R_j^{(1)}$, $j = 1, \ldots, N_1$ and $R_j^{(2)}$, $j = 1, \ldots, N_2$ from the first and second set of contours, respectively. Averaged radii R_j, taken as random variables, are independent (they are computed at different objects). Moreover, in the population of one type of objects, they are assumed to be identically distributed. That is why the comparison of averaged radii of both types of cross-sections can be accomplished with the aid of a simple two-sample statistical test.

2.1. **The test for nonequal radii of shapes.** The type of distribution of R_j is not known, though we could use a proper version of the central limit theorem and assume approximate normality. We have to take into account also the fact that as the points of contours were selected rather close one to each other, r_i for neighboring i-s (at the same contour) were mutually dependent, so that R_j were averages from mutually dependent variables. We shall return to that problem later, when dealing with the regression model. At the present moment, this problem can be overcome by the use of one from nonparametric two-sample tests, instead of the standard t-test.

Let us assume that the values $R_j^{(k)}$, $k = 1, 2$, $j = 1, \ldots, N_k$, represent two random variables $R^{(1)}$, $R^{(2)}$, resp., and let us consider the hypothesis H_0 that $R^{(1)}$ does not differ systematically from $R^{(2)}$, against the one-sided alternative $H_1: R^{(1)} < R^{(2)}$. For instance two-sample test of Wilcoxon based on the order statistics can be applied. However, as all values $R_j^{(2)}$ are greater than all $R_j^{(1)}$ (it is also shown by the histogram of values $R_j^{(k)}$ on Figure 2), the rejection of H_0 in favour of H_1 is evident.

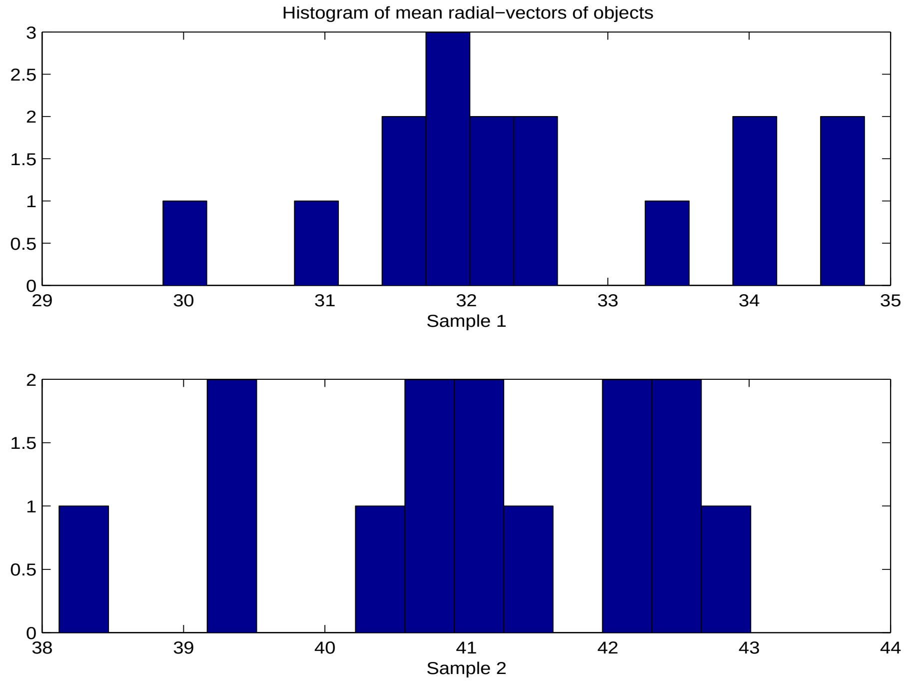


Fig. 2 *Histogram of mean radii of objects.*

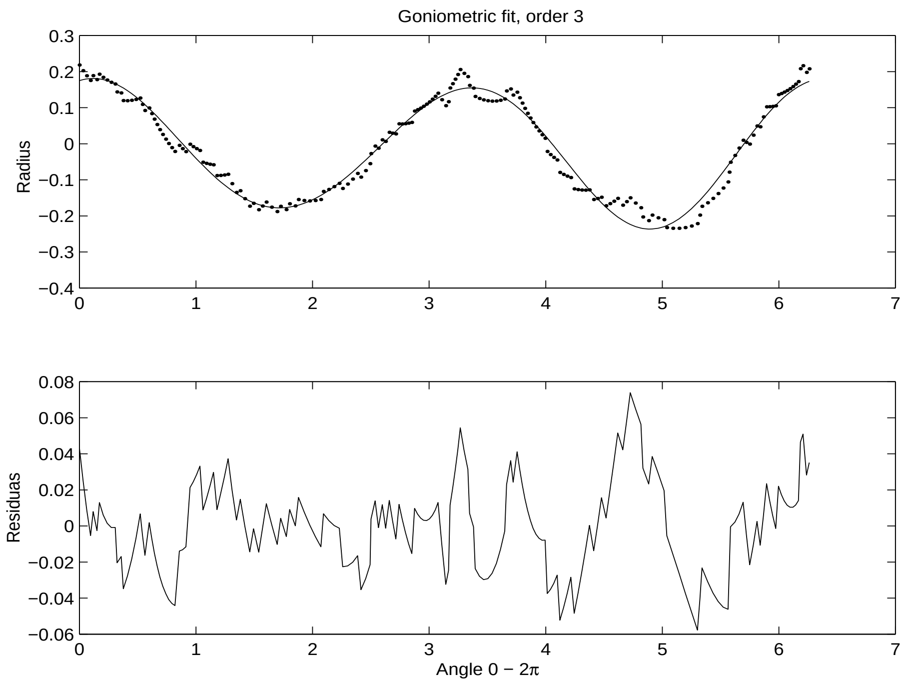


Fig. 3 *Example of unfolded contour data, with optimal model function.*

3. Regression model of unfolded contours

Let us first remove the influence of different lengths of radii, of each contour j, by the transformation $v_i = r_i/R_j - 1$. Each contour is now represented by the data $\{(u_i, v_i),\ i = 1, 2, \ldots, n_j\}$, where u_i are angles from 0 to 2π and v_i are 'normalized' radii. Figure 3 (upper subplot, the point-wise curve) shows the example of unfolded contour of one cross-section. It is seen that the curve has a periodic character – similar to an unfolded ellipse with additional nonregularities. Therefore it is expected that the combination of trigonometric functions will provide a good functional model in this case. In the framework of nonlinear regression model $v_i = g(u_i) + \varepsilon_i$, let us, for each ($j$-th) unfolded contour, consider the following linear model (i.e. linear with respect to parameters):

$$v_{ij} = a_{0j} + \sum_{k=1}^{K} (a_{kj} \sin(k u_{ij}) + b_{kj} \cos(k u_{ij})) + \varepsilon_{ij}. \tag{1}$$

Such a model is quite frequently used in signal analysis, where the trigonometric function is often combined with a trend function (linear or quadratic, for instance). However, as it has already been noted, the points of contours were selected rather close one to each other in original 'densely' sampled data, so that ε_{ij} for neighboring i's (and the same j) were mutually dependent. The additional statistical analysis revealed that this dependence can be well modeled as a linear autoregression of order 1 or 2. Optimal order of AR model has been determined with the aid of Schwarz BIC criterion, a standard criterion used in this field of statistical methodology. However, when the data were reduced, namely when only each 5-th point (u_i, v_i) of original contour data was taken, the mutual dependence of neighboring points disappeared practically. Namely, the regression model (1) was applied to the reduced data (now the sample size of data for one contour was between 40 and 50). Then the residuas estimating the departures ε_{ij} were computed, and their mutual independence tested by tests of randomness. More precisely, 'runs up and down' and 'runs above and below the median' tests were used and the hypothesis of randomness was not rejected, as a rule by any of tests.

Therefore, the regression model (1) has been fitted to reduced data. Then ε_{ij} were already regarded as mutually independent, centered, symmetric random variables, identically distributed at least for each j. The parameters of the model were estimated in a standard way, by the least squares method, which was accompanied by the estimate of residual variance (of variables-noises ε_{ij})

$$\hat{\sigma}_j^2 = \sum_{i=1}^{n_j} (v_{ij} - \hat{g}_j(u_{ij}))^2 / (n_j - 2K - 1).$$

We should discuss also a rather important problem of optimal model complexity (i. e. of optimal selection of K). One way indicating the non-significance of certain parameters a_k, b_k can be based on the standard t-tests testing the hypothesis $a_k = 0$ (resp. $b_k = 0$) separately for each parameter. We actually use the normal approximation instead of the t-test, because, though we do not assume the normality of noises ε_i, on the other side we deal with rather large sample sizes n_j's, so that the normal approximation is adequate. However, the Schwarz 'BIC' criterion was again used as the main (though ad-hoc) criterion of optimal selection of K. Namely, we selected such a model that $\ln \hat{\sigma_j}^2 - 2K \ln n_j / n_j$ was minimal (where $\hat{\sigma_j}^2$ was the estimate of residual variance). In the most cases, the optimal model had $K = 3$. In several instances, moreover, the t-tests denoted some coefficients as nonsignificant so that the model could be further reduced.

Figure 3 shows one example of such a regression curve of order 3 (full line in the upper plot) and also the sequence of residual values (lower plot). As the data were shifted in such a way that the curve started from its maximum, even the models using only the cosinus functions were quite good (mostly with optimal order $K = 5$).

		Type 1:			Type 2:		
	Parameter	Mean	STD	Median	Mean	STD	Median
1	a_0	-0.0126	0.0097	-0.0098	-0.0078	0.0061	-0.0063
2	a_1	0.0002	0.0053	-0.0001	-0.0032	0.0119	-0.0027
3	b_1	-0.0181	0.0221	-0.0131	-0.0115	0.0087	-0.0099
4	a_2	-0.0034	0.0603	-0.0061	-0.0147	0.0540	-0.0102
5	b_2	-0.1442	0.0550	-0.1385	-0.0918	0.0558	-0.0855
6	a_3	0.0010	0.0162	0.0034	0.0019	0.0170	0.0078
7	b_3	-0.0352	0.0308	-0.0330	-0.0253	0.0148	-0.0264

Tab. 1 *Means, standard deviations and medians of parameters.*

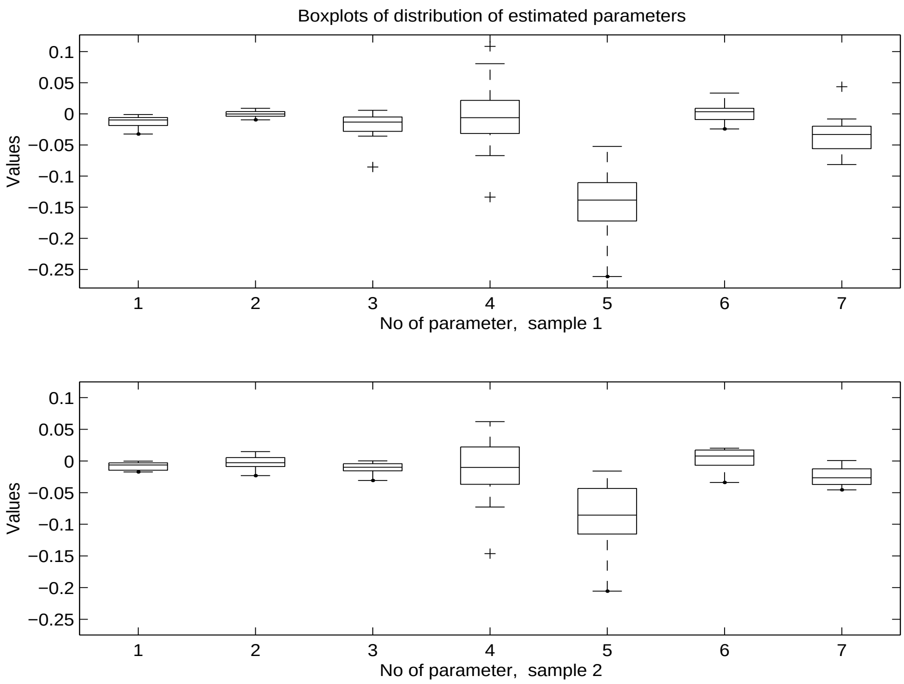


Fig. 4 *Boxplots of estimated parameters of model (1).*

Thus, the results of the procedure described above were N_1 sets of seven parameters $(a_0, a_1, b_1, a_2, b_2, a_3, b_3)$ obtained (estimated) from the first sample of cross-sections, and N_2 sets from the second sample. Table 1 displays the mean values, standard deviations, and medians of these sets. Estimated characteristics of distributions of parameters in both sets were compared. Graphical comparison is seen from the boxplots of Figure 4. The numerical comparison was again performed with the aid of 2 sample t-tests. The tests lead to the conclusion that the only significant difference was the difference between parameters b_2 (the statistics $T = -2.5860$), the mean value of b_2 from the sample 1 was -0.1442 against -0.0918 from the sample 2. Parameter b_2 corresponded to the component $\cos(2u)$, which was the component most influencing the deformation of the contour. Therefore we might conclude that the heat-treated fibers had significantly flattened cross-sections, compared to the sample of heat-untreated fibers.

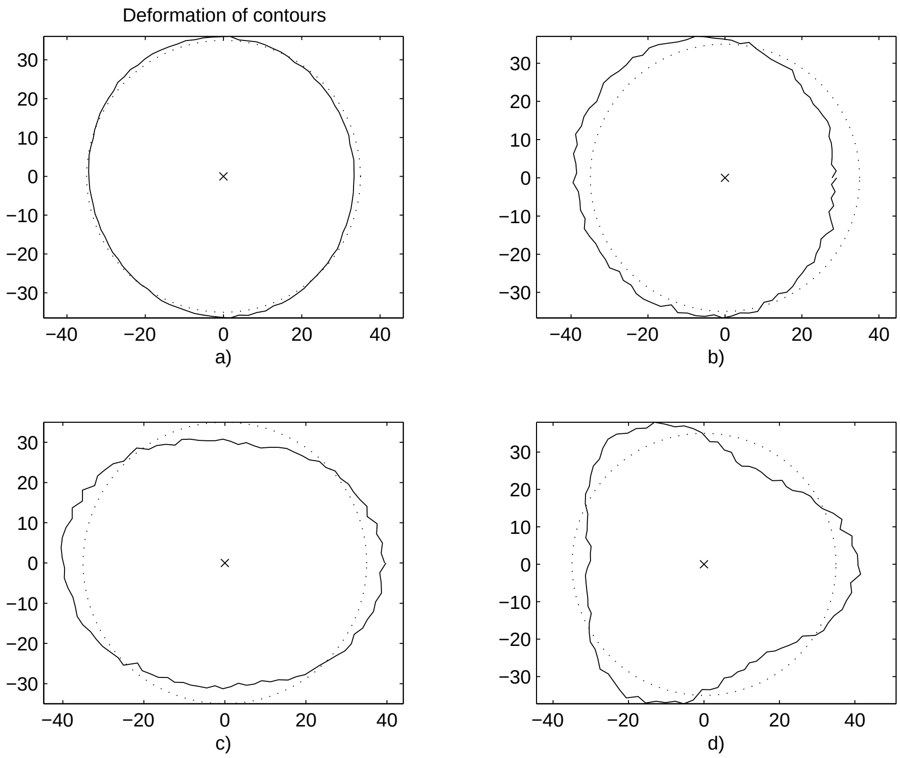


Fig. 5 *Simulated shapes.*

4. Simulation of contours

The objective of the simulation was to support the conclusion of statistical analysis, by random generation of contours using the model (1). One set of results is shown in the Figure 5. We selected 100 equidistant points u_i between 0 and 2π and generated corresponding normalized radii v_i in accordance with the model (1). Then, the random noise was added; it was generated from the normal distribution with zero mean and the variance corresponding to averaged residual variances obtained from our data ($\sim 4 \cdot 10^{-4}$). Finally, the resulting function with random noise was added to a regular circle with radius 1 and then multiplied by 35, so that the obtained contour represented a 'noisy' deformed circle with radius ~ 35. The first case displayed in Figure 1 corresponds to the circular contour without deformation, i.e. the case when all parameters of model (1) were set to zero, $g(u_i) \equiv 0$. The result is on subplot a). Then, one of the parameters was changed, while the rest was still kept equal to zero. Thus, subplot b) corresponds to the case with decreased $b_1 = -0.1$ (the contour is shifted). Subplot c) displays the result of decreased $b_2 = -0.1$ – the contour is flattened, which corresponds to the situation detected in our real case data. Finally, subplot d) shows the result of decreased parameter $b_3 = -0.1$. The changes of parameters a_1, a_2, a_3, respectively, had the same consequences (only the direction of deformation varied). Thus, the results of simulations supported the conclusion that the changes of certain parameters are connected with certain types of deformation.

On the other side, parameters b_2 of model (1) do not suffice to a direct classification of shapes. Figure 5 shows that there is not any distinct border between the parameters of the first and the second group and that their distributions overlap. An experiment with the classification tree construction showed that a combination of at least three parameters is necessary for a good separation of both groups of contours.

5. Concluding remarks, Procrust analysis

From the point of view of mathematical shape analysis, the proposed model was a rather simple one. Nevertheless, it quite sufficed for the conclusion that one type of contours differs significantly from the other. We also demonstrated how such a model was able to describe certain kinds of deformations of circular shapes.

We want to mention here also another approach to the shape analysis, namely the Procrust analysis (e.g. Dryden and Mardia, 1998). The shape is represented by a set of significant points (landmarks) placed on the contour. The relative positions of these landmarks are then compared (either with landmarks on a standard object or among different objects). The typical applications are the recognition of certain objects from the images of earth surface or the analysis of CTM images in medical studies. On the other hand, the method is not convenient for shapes without natural significant points, or for the shapes too complicated, when a large number of points is necessary for shape description. The case of circular shapes (i.e. the case considered in our study) belongs to the first group. For such instances the method can be enriched by the analysis of cyclic permutations of selected contour points. However, neither such a modification was efficient enough to solve our problem, i.e. to discriminate clearly between examined samples of fibers.

Reference

[1] A. J. Baddeley and M. N. M. van Lieshout: Stochastic geometry models in high-level vision. *In Statistics and Images, Advances in Applied Statistics, K. V. Mardia and G. K. Kanji ed., Suppl. to the Journal of Applied Statistics* **20**, 231–261. Carfax Publishing, 1993.
[2] N. A. C. Cressie: Statistics for Spatial Data. Wiley. New York, 1995.
[3] I. L. Dryden and K. V. Mardia: Statistical Shape Analysis. Wiley, New York, 1998.
[4] U. Grenander: General Pattern Theory. Oxford University Press, New York, 1993.
[5] A. Hobolth, J. Pedersen, and E. B. V. Jensen: A continuous parametric shape model. Research Report N.13, Lab. for Computational Stochastics, University of Aarhus, 2000.
[6] H. Rue and M. A. Hurn: Bayesian object recognition. *Biometrika*, **86**, 649–660, 1999.

KTM TU Liberec, Hálkova 6, 461 17 Liberec
E-mail: ales.linka@vslib.cz

ÚTIA AV ČR, Pod Vodárenskou věží 4,182 08 Praha 8
E-mail: volf@utia.cas.cz