

Interpretable graph methods for determining nanoparticles ordering in electron microscopy image

M.Y. Kurbakov¹, V.V. Sulimova¹, O.S. Seredin¹, A.V. Kopylov¹

¹ Tula State University, 300012, Russia, Tula, Lenina Av. 92

Abstract

An important step in determining the properties of carbon materials is the analysis of images from a scanning electron microscope (SEM). These images show the material surface after the application of metal nanoparticles. The order of these nanoparticles is a key characteristic that affects the material properties. We have previously proposed an approach to formalize the order features based on the identification of lines by nanoparticles in the SEM image. This paper proposes a novel approach to line allocation that is based on the concept of constructing a minimum spanning forest. Additionally, it introduces a set of novel ordering functions that are derived from this approach. The experimental study demonstrates that the combination of these new and previously extracted features improves the recognition quality of SEM images with ordered and disordered nanoparticles arrangements. This approach allows us to gain a better understanding of the nanoparticles arrangement and their effect on the material properties.

Keywords: explainable machine learning, image analysis, nanoparticle detection, nanoparticles ordering features.

Citation: Kurbakov MY, Sulimova VV, Seredin OS, Kopylov AV. Interpretable graph methods for determining nanoparticles ordering in electron microscopy images. *Computer Optics* 2025; 49(3): 470-479. DOI: 10.18287/2412-6179-CO-1568.

Acknowledgements: This work was supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of the state task FEWG-2024-0001.

Experiments were partially made using the equipment of the shared research facilities of HPC computing resources at the Lomonosov Moscow State University.

The authors thank the Scientific School of Academic V.P. Ananikov for the research topic, useful discussions and provided experimental data.

Introduction

Carbon-based materials are essential in modern materials science [1], with applications in the fields of electronics, sensors, adsorption, purification, and catalysis. The specific properties of carbon materials depend heavily on their structure [2], making it critical to determine this structure for optimal use.

Scanning electron microscopy (SEM) is a widely used method for studying these materials [3, 4]. However, it can be challenging to study some surface features of materials using this method, such as defects. To address this challenge, researchers have developed modern approaches based on the use of palladium metal nanoparticles [5]. These nanoparticles are applied to the surface of a material and, due to their unique electro-chemical properties, they can act as markers for defects.

In recent years, computational modeling, in general, and machine learning algorithms in particular, have been actively employed in nanotechnology [6–8]. However, the analysis of electron microscopy images, which provides insights into the arrangement of particles, remains a significant challenge. While much work has focused on detecting individual particles, research on more global material features, especially those related to materials ordering, is limited, especially when considering human analysis [9]. A major target for analysis is the distinction between ordered and disordered materials, and the determination of their overall order.

Generally, deep learning methods [9] are effective tools for discovering dependencies within experimental data. However, these methods do not provide an explanation for the rationale behind these dependencies, storing the acquired knowledge in the weights and parameters of a trained model. This representation is insufficient for scientific investigations, where an explanation of the dependencies is crucial.

In our previous work [10], we provided researchers with understandable and controllable features that can lead to a reliable interpretation of the underlying chemical processes. We did this by using visualization data from metallic nanoparticles and by formalizing the concept of an ordered relative arrangement of these nanoparticles, considering the specifics of human perception. As a result, we formed a number of features based on the linear structures of nanoparticles, which were obtained using the principal components method and the shortest unclosed path.

This article proposes a novel method for extracting linear structures by constructing a minimum spanning forest and forming a new set of ordered features based on this approach. These features will be used in the future for explanatory analysis of the ordering phenomenon.

1. Suppositions about nanoparticles ordering

The central notion of this work is the notion of nanoparticles ordering. We assume that nanoparticles are or-

dered if they form lines. At that the more lines and the longer and "smooth" they are, the more ordered the nanoparticles arrangement is.

In accordance with this supposition to estimate nanoparticles ordering we construct lines of nanoparticles. It is important, that within the framework of this paper the initial data for lines construction are the only coordinates of nanoparticle centers, that were obtained at the previous study in accordance with the exponential approximation method proposed in our work [11]. So, in contrast to [9] we do not take into account any regular structures of material surface that can be seen at the SEM images.

It should be noted that the problem under consideration has an important feature that significantly complicates its automatic solution. This feature consists in the presence of (quite typical) cases when the traditional Euclidean distance between nanoparticles belonging to neighboring lines distinguished by human vision can be comparable or even less than the Euclidean distance between nanoparticles of one line.

In this regard, to determine the neighborhood of nanoparticles while lines construction, a new specialized metric is applied here, which was initially proposed in our work [10] and is described in details at the subsection 2.3.

To construct lines of nanoparticles we use two different approaches that are differ by a type of binding nanoparticles:

- 1) on the basis of the shortest unclosed path method [10] (subsection 3.1)
- 2) a new approach on the basis of a minimum spanning forest (subsection 3.2).

Section 4 proposes a small number of highly interpretable features extracted from found lines of nanoparticles. These features are interpretable and informative for making an automatic decision about the presence or absence of order in a nanoparticles arrangement. At the same time, the results of the experimental study (subsection 5.2) show that joint use of all features (proposed in our work [10] and in this paper) allows to reach the best recognition quality.

2. Prevailing directions metric

The prevailing direction of a local nanoparticles group is defined as the direction in which the nanoparticles of this group line up.

The prevailing directions metric in addition to the Euclidean distances between points (nanoparticles centers), takes into account the prevailing directions of nanoparticles local groups and their reliabilities.

This metric was initially proposed and fully described in our work [10]. This section contains main formal definitions of the proposed approach, such as a nanoparticles local group, the prevailing direction of a nanoparticles group and its reliability and then introduces the prevailing directions metric based on them.

2.1. Local groups of nanoparticles

Forming any local group of nanoparticles is started from one nanoparticle. A new nanoparticle for adding to a group is selected as a nanoparticle with the minimum Euclidean distance to the nearest nanoparticle of the group in accordance with the modified Prim's algorithm [12]. Our proposed modification [10] consists in applying early stopping criterion, that allows to stop forming a local group before reaching the maximum local group size s . Early stopping is realized on the basis of the threshold that has the meaning of the average local nanoparticles density d in areas with their most intense accumulation and can be estimated on the basis of $k \cdot N$ minimal distances between nanoparticles $e_i \in E$, $i = 1, \dots, k \cdot N$, where $E = [e_{ij}; i, j = 1, \dots, N]$ is the matrix of Euclidean distances between all nanoparticles.

$$d = w_d / (k \cdot N) \cdot \sum_{i=1}^{k \cdot N} e_i, \quad (1)$$

where N is the number of nanoparticles in the SEM image, k is the proportionality coefficient, that defines the number of used minimal distances and w_d is a weight coefficient.

2.2. The prevailing direction of a local group and its reliability

The prevailing direction of a local nanoparticles group (as the direction along which the nanoparticles of this group line up) corresponds to the maximum eigenvector of the covariance matrix of a pair of vectors that are composed of the nanoparticles centers coordinates. The tilt angle of this vector θ relative to the horizontal can be calculated by the formula [13, 14]:

$$\theta = \arctan(2\mu'_{11} / (\mu'_{20} - \mu'_{02})), \quad (2)$$

where μ'_{11} , μ'_{20} , μ'_{02} – elements of the covariance matrix of a pair of vectors, that are composed of the coordinates of the centers of nanoparticles.

To estimate the reliability of the prevailing direction, the eccentricity [14] is used:

$$q = (1 - \lambda_{min} / \lambda_{max})^2, \quad (3)$$

where λ_{max} и λ_{min} are the maximum and minimum eigenvalues of the covariance matrix, respectively.

This estimate takes values in the range [0, 1] and shows how much the arrangement of nanoparticles of the local group is "elongated" along the prevailing direction. At that the best possible value $q = 1$ is reached in the case, when all nanoparticles of the local group are located on the same straight line.

2.3. Prevailing directions metric definition

The proposed metric of prevailing directions (PD) is constructed as the weighted combination of the Euclidean distance and the prevailing directions with their corresponding reliabilities. Then the distance m_{ij} between nanoparticles i and j can be calculated using the formula:

$$m_{ij} = C \cdot e_{ij} + 2 \cdot (1 - C) \times \max \left[\sin \left(\left| \theta_i - \theta_j \right| \right), \left(\varepsilon_i + \varepsilon_j \right) / 2 \right], \quad (4)$$

where e_{ij} is the Euclidean distance between nanoparticles centers i and j ; θ_i is the prevailing direction tilt angle for the local group corresponding to the i -th nanoparticle (2); ε_i is unreliability of the corresponding i -th prevailing direction, which is the reciprocal of the reliability q_i (3), $\varepsilon_i = 1 - q_i \in [0, 1]$; C is the proportionality coefficient, that allows us to adjust the influence degree of individual metric parts on its resulting value.

Note that the difference in the tilt angles of the prevailing directions can be characterized by one of the adjacent angles at the intersection of these directions. Since it does not matter which of the neighboring angles is used for the sine, we use the sine of the angle difference of the inclination directions to estimate the difference between them. Also, the use of a sine allows us to normalize the magnitude of the angle difference.

As a result, according to the proposed metric, the higher the average prevailing directions unreliability is, the higher metric value is (the respective nanoparticles are more distant from each other). At the same time, if the difference in tilt angles is large, then the corresponding nanoparticles will be considered distant even for small average unreliability due to the presence of the maximum operation.

3. Nanoparticles lines construction

It is important, that within the framework of this paper the initial data for lines construction are the only coordinates of nanoparticle centers.

Fig. 1 shows examples of two typical SEM images from the base [18] with ordered (Fig. 1a) and disordered (Fig. 1b) nanoparticles arrangement, and the results of finding nanoparticles for each of them without a substrate material (Fig. 1c, d).

3.1. Lines on the basis of the shortest unclosed path method

In this case each line is formed point by point. An initial point of a line is selected from a set of nanoparticles that have not been connected in lines yet as a nanoparticle with the maximum reliability (3) of the prevailing direction of the local group (subsection 2.1) associated with it. Such choice corresponds to the maximum "elongation" of nanoparticles of the corresponding local group along the prevailing direction.

To ensure line smoothness we additionally dynamically correct metric values immediately in process of line construction. Then the adjusted distance m_{ij}^* between nanoparticles i and j can be calculated using the formula:

$$m_{ij}^* = m_{ij} + w_{coax} \cdot (1 - coax_{ijk}^2), \quad (5)$$

where m_{ij} is the basic part of the prevailing directions metric (4), $coax_{ijk} = (1 - \cos \alpha) / 2 \in [0, 1]$ is so called an-

gular coaxiality, that measures the difference α between tilt angles of the previous line segment (i, j) and the potential next segment (j, k), that can be added to the line and w_{coax} is a weight coefficient of the angular coaxiality.

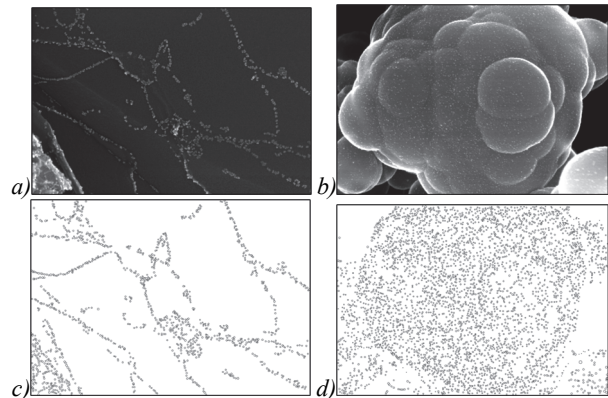


Fig. 1. Examples of two typical SEM images with nanoparticles arrangement: (a) ordered, (b) disordered.

(c), (d) the results of finding nanoparticles for each of them without a substrate material

Each next point to add to the line can be found on the basis of the shortest unclosed path (SUP) method [15], which we modified to incorporate dynamically computed distances (5), to elongate the line in two sides (by adding new points before the first and after the end point) and to stop line forming if the corrected PD distance (5) exceeds the adopted threshold thr , that is computed as a special case of (4):

$$thr = C \cdot d + (1 - C) \cdot w_{thr}, \quad (6)$$

where d is the nanoparticles density, estimated for a SEM image in accordance with (1); $w_{thr} = 1$ is the value, that is obtained for the limiting values of the angle θ and unreliability ε in (4); C is the same proportionality coefficient of PD metric as in (4).

More detailed description of the proposed metric properties and of the modified SUP algorithm are provided in our work [10].

3.2. Lines on the basis of a minimum spanning forest

This section proposes a new way of binding nanoparticles in lines, that is based on constructing a minimum spanning forest on a set of nanoparticles detected in a SEM-image.

As, mentioned above, each nanoparticle in this work is described by coordinates of its center and a radius. This information allows us to easy compute a distance between any two nanoparticles (it should be noted, in this study we consider distances between centers of nanoparticles, ignoring their sizes).

As a result, a set of nanoparticles can be considered as undirected fully connected cyclic edge-weighted graph, whose nodes are nanoparticles. Each edge of this graph connects two nanoparticles and has the weight, which is equal to distance between them.

In this paper we propose to find lines of nanoparticles via a minimum spanning tree (MST) graph-theoretic technique which determines a subset of the edges of such graph that connects all the nodes together, without any cycles and with the minimum possible total edge weights [16]. So, in our case MST allows to determine shortest path of nearest-neighbor connections of nanoparticles in terms of some distance.

At the same time, it should be noted, that from the point of view of the considered applied problem, distant nanoparticles should not be connected in lines and so, we reduce the initial graph by removing those edges from it, whose length is greater than a certain threshold $dist_{max}$, that is the parameter of the proposed method.

As a result, the initial graph generally splits into unconnected parts and applying the MST technique for it is resulted in a minimum spanning forest (MSF), which is a union of the minimum spanning trees for its connected components. To construct minimum spanning forest, we apply the Kruskal's algorithm [17], that starts construction from minimal distances and incrementally adds new longer edges to the graph if they don't form a cycle.

The nodes of obtained graph can have a different degree of branching. The branching degree of some node determines the number of edges emerging from it. So, if a node has a degree equal to zero, then this means that the corresponding nanoparticle is not connected to any other nanoparticle. Nodes with branching degree greater than two will be called branch points.

Then, by a line connecting nanoparticles, we will mean here a sequence of edges located between two branch points or between a branch point and a terminal node (having a branching degree equal to one). However, it should be noted that the obtained graph as a rule has multiple one-edge offshoots. They form multiple additional branching points and, as a result, interfere with the long lines detection in the case of ordered nanoparticles.

In this connection to improve distinguishing between ordered and disordered arrangement of nanoparticles we propose make an additional processing here. Immediately, we propose to remove from the graph one-edge lines that immediately connect a branch point and a terminal node. Such lines we name terminal one-edge offshoots. As a result of this processing the obtained graph has smaller number of branching points and its lines are longer.

Fig. 2 shows examples of minimum spanning forest for nanoparticles before (Fig. 2a) and after (Fig. 2b) removing one-edge terminal offshoots (for the SEM image fragment on Fig. 1c). Blue indicates those lines whose length is greater than or equal to 15 nanoparticles.

It should be noted, that the proposed approach is essentially based on a distance measure between nanoparticles. At that it is evident, the different kind of distance will lead to different minimum spanning forests and, respectively, different lines resulted from them.

In this paper, along with the traditional Euclidean metric, we use the prevailing directions metric (4) that

was originally proposed in our work [10], which in a number of cases makes it possible to obtain smoother and longer lines in contrast to the Euclidean metric.

Fig. 3 shows the lines found using the minimum spanning forest approach using the Euclidean metric (Fig. 3a, b) and the Prevailing Directions metric (Fig. 3c, d) for two typical SEM images with ordered (Fig. 1c) and disordered (Fig. 1d) nanoparticles arrangement. Fig. 4 shows for comparison the lines constructed using the method based on finding the shortest unclosed path [10].

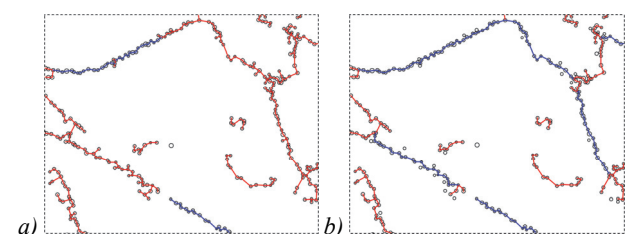


Fig. 2. Examples of minimum spanning forest for removing one-edge terminal offshoots: (a) before, (b) after. Blue lines are those whose length is greater than or equal to 15 nanoparticles

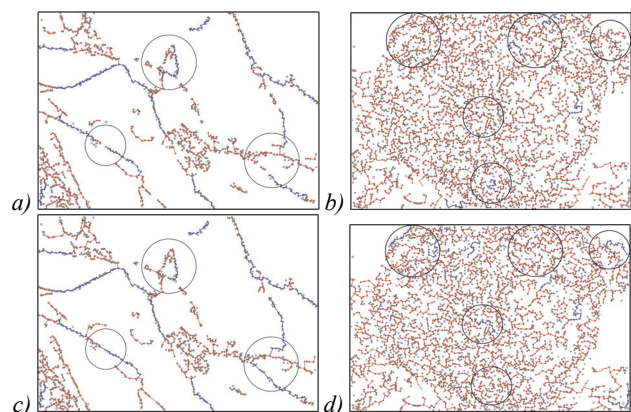


Fig. 3. The lines found based on the minimum spanning forest method for two typical SEM images using: (a), (b) the Euclidean metric; (c), (d) the metrics of Prevailing Directions (PD). Blue lines are those whose length is greater than or equal to 15 nanoparticles. Big circles indicate areas of the most significant difference of minimum spanning forests with different metrics

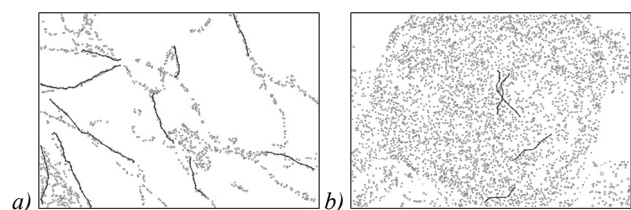


Fig. 4. The lines found based on the shortest unclosed path method for two typical SEM images with nanoparticles arrangement: (a) ordered, (b) disordered. The length of the lines is greater than or equal to 15 nanoparticles

As we can see from the Fig. 3, each of the proposed approaches has a tendency to find more lines in the case of ordered nanoparticles in contrast to disordered case and, at the same time, the found lines, as a rule, are more long and smooth. As expected, the smoothness is espe-

cially clear seen in the case of using the modified shortest unclosed path method with the corrected dynamically computed PD metric (5).

4. Interpretable features to determine nanoparticles ordering

4.1. Features on the basis of prevailing directions (O-features)

The features O_1 and O_2 are based on the assumption that the ordered arrangement of nanoparticles is characterized by the presence of a significant number of local groups with similar (or even identical) prevailing directions (such directions are named consistent ones here), which is due to the nanoparticles of corresponding groups are arranged along long smooth lines. At the same time for the disordered nanoparticles arrangement, significantly more uniform distribution of prevailing directions is typical. So, in the case of the disordered arrangement weak directions consistency takes place.

A quantitative measure of general consistency of prevailing directions can be computed on the basis of the Shannon entropy [19]. Tilt angle of the prevailing direction always takes values in the range $[-90^\circ, +90^\circ]$. To calculate the Shannon entropy this range is divided into m intervals of the equal length and empirical probabilities p_i of the angle falling into each interval $i = 1, \dots, m$ are computed. Then the value of the Shannon entropy H can be calculated by the formula:

$$H = -\sum_{i=1}^m p_i \cdot \log_2 p_i. \quad (7)$$

Note, that the maximum possible entropy value is limited and can be reached in the case of a uniform distribution [20]:

$$H^* = \log_2 m. \quad (8)$$

The final value of the O_1 feature can be calculated as the ratio (7) on (8) and reflects the effective value [20]. The negative sign in the ratio is required to normalize the values of the features – the greater, the better.

$$O_1 = -H / H^*. \quad (9)$$

In this case $O_1 \in [-1, 0]$ and the highest value of 0 can be achieved when all nanoparticles are arranged in a single line. The value of m in this work was taken as 90.

The feature O_2 is a special case of the general prevailing directions consistency. The difference consists in that only tilt angles of those prevailing directions are taken into account, whose reliability (3) exceeds the specified threshold q_{min} .

Each prevailing direction is characterized by its reliability (3), which form the set of reliability values $Q = \{q_1, \dots, q_N\}$, where N is the number of detected nanoparticles. Also, we consider a subset of reliable directions, whose reliability values exceed the specified threshold: $Q^* = \{q_1^*, \dots, q_n^*\} \subseteq Q$, $q_i^* > q_{min}$, $i = 1, \dots, n$.

The proposed O_3 feature defines the proportion of reliable directions taking into account the reliability values oneself at that:

$$O_3 = \sum_{i=1}^n q_i^* / \sum_{i=1}^N q_i. \quad (10)$$

In this case $O_3 \in [0, 1]$, where the larger its value, the more ordered nanoparticles arrangement is.

4.2. Features on the basis of nanoparticle lines found by the shortest unclosed path method (L-features)

The L_1 feature is the number of lines, constructed on the basis of the modified SUP method, whose length in nanoparticles is greater or equal to a threshold L_{min} .

The L_2 and L_3 features are numerical estimates of the rectilinearity and the smoothness of the constructed lines, that directly correspond to the human perception of the fact of orderliness. To formalize these notions, we introduce a special characteristic named a metric coaxiality of a polyline.

Let $P_1^n = p_1, \dots, p_n$ be a constructed polyline consisting of n ordered nanoparticles. Then the metric coaxiality for this polyline can be defined as follows:

$$coax(P_1^n) = e(p_1, p_n) / \sum_{i=1}^n e(p_i, p_{i+1}), \quad (11)$$

where $e(x, y)$ is the Euclidean distance between centers of nanoparticles x and y .

The rectilinearity of a polyline P_1^n expresses its similarity to a straight line throughout it and can be calculated as the metric coaxiality (11) of the full polyline:

$$R(P_1^n) = coax(P_1^n). \quad (12)$$

The L_2 feature characterizes smoothness of all constructed lines at once and is computed by averaging individual rectilinearity values (12).

The smoothness of a polyline P_1^n expresses its local similarity to a straight line and can be computed as averaged metric coaxiality (11) of all polyline fragments $P_a^b = p_a, \dots, p_b \in P_1^n$ of some size $f_{size} = b - a + 1$:

$$S(P_1^n) = 1 / m \cdot \sum_{i=1}^m coax(P_i^{f_{size}}), \quad (13)$$

where $m = n - f_{size} + 1$ is the number of fragments of size f_{size} on polyline P_1^n .

The L_3 feature characterizes smoothness of all constructed lines at once and is computed by averaging individual smoothness values (13).

The L_4 feature show proportion of connected on lines nanoparticles and calculated as the ratio of the number of nanoparticles that are connected by all the constructed lines to the total number of detected nanoparticles.

4.3. Features on the basis of nanoparticle lines found by the minimum spanning forest approach (B-features)

To characterize each image on the basis of the Minimum Spanning Forest (MSF) we compute next characteristics (based on the lines from branches).

Features, similar to L_3 and L_3 , are also formed for MSF-based lines. The B_1 feature is the mean rectilinearity (12) of lines in the MSF graph. The B_2 feature is the mean smoothness (13) of lines in the MSF graph. But they are calculated only for the N_l longest lines (length is determined by the nanoparticles number in line). If the total number of formed lines is less than N_l , then these features are calculated for all constructed lines.

The B_3 feature is the proportion of nanoparticles belonging to lines that are at least N_n nanoparticles long of the total number of nanoparticles in the SEM-image.

The B_4 feature is the mean length of N_l longest lines in the MSF graph, normalized to the maximum line length in nanoparticles.

It should be noted, each of the indicated feature values can be computed for lines on the basis of the Euclidean metric and on the basis of the proposed metric of prevailing directions (4). Next, we will identify the features calculated based on the Euclidean distance as B^E , and based on the metric of Prevailing Directions as B^{PD} .

5. Experiment's description

5.1. Data collection

A previously published dataset [18] with 750 images with a particle ordering effect and 250 images without an ordering effect was used. SEM images are in the TIFF format with 1280×890 resolution. The images are already separated into two groups: with predominantly ordering and with predominantly disordering effects. Table 1 shows the SEM images distribution by material types and scales.

Tab. 1. SEM images distribution by material types and scales

		Material with effect					Sum
		ordering		disordering			
		S1	S2	S3	S4	S5	
Scale	50k	203	34	63	25	4	329
	100k	322	29	63	25	11	450
	200k	162	-	59	-	-	221
	Sum	687	63	185	50	15	1000

It has previously been noted that the proposed method relies solely on the coordinates of nanoparticle centers, and therefore, the results of nanoparticle detection serve as input data. In this work, the previously proposed exponential approximation method is used to detect nanoparticles. For most of the images in the dataset, the exponential approximation method default parameters specified in our work [11] are adequate. However, some images differ from the rest in terms of their average brightness, which can be significantly higher or lower. In such cases, the selection of parameters for the exponential approximation method must be carried out individually for each image. For more information about the parameters of the exponential approximation method see [11].

5.2. Determining nanoparticles ordering

In previous work [10], we showed that a linear SVM classifier in the feature space based on the shortest un-

closed path (L -features) and prevailing directions (O -features) demonstrates classification quality comparable to the neural network approach [9]. The classifier quality was estimated using a 5-fold cross-validation [21] procedure with stratification [22] (since the data set is unbalanced: 750 ordered and 250 disordered images).

The main classification quality indicators are calculated based on the confusion matrix of classification, which contains the following values: TP - orrectly classified ordered images, TN -correctly classified disordered images, FP -ordered images classified as disordered, and FN - disordered images classified as ordered. Then the classification accuracy (Acc) [23] is the ratio of correctly recognized images to the total number of images:

$$Acc = (TP + TN) / (TP + FN + TN + FP). \quad (14)$$

At the same time, most classifiers can balance the decision rule either toward increasing the number of correctly recognized positive class objects (ordered) or toward reducing the number of incorrectly classified negative class objects (disordered) using some hyperparameters. In this regard, such characteristics as *Precision*,

$$Precision = TP / (TP + FP), \quad (15)$$

which is understood as the proportion of correctly recognized objects among all objects recognized as positive, and *Recall*,

$$Recall = TP / (TP + FN), \quad (16)$$

which represents the proportion of correctly recognized objects among all positive objects, are often used.

The F-measure (F) is a widely known measure that attempts to combine these two indicators and characterize the quality of the classifier with a single number [23]. It is defined as the harmonic mean between *Precision* and *Recall*:

$$F = 2 \cdot Recall \cdot Precision / (Recall + Precision). \quad (17)$$

P_4 [24] is a variant of the F-measure that does not depend on which class is called positive and which is negative, and is able to produce accurate results with unbalanced data. Thus, the quality of the classifier will be evaluated by the four elements of the original confusion matrix:

$$P_4 = 4 / \left[(1/TP + 1/TN) \cdot (FP + FN) + 4 \right]. \quad (18)$$

Tables 2 show the main quality indicators on 5-fold cross-validation for the linear SVM classifier based on the O - and L -features, as reported in our work [10].

Tab. 2. Linear SVM classifier quality indicators for all O - and L -features

Target class	Acc	Precision	Recall	F	P_4
Ordered	0.950	0.967	0.967	0.967	0.932
Disordered		0.902	0.900	0.899	

The addition of features based on the minimum spanning forest (calculated using the Euclidean metric - B^E and the metric of Prevailing Directions - B^{PD}) makes it possible to significantly improve the quality of classification. Tables 3 show the main quality indicators on 5-fold cross-validation for the linear SVM classifier based on the O -, L - and B^E -, B^{PD} -features.

Tab. 3. Linear SVM classifier quality indicators for all O -, L - and B^E -, B^{PD} -features

Target class	Acc	Precision	Recall	F	P_4
Ordered	0.967	0.980	0.976	0.977	0.955
Disordered		0.930	0.940	0.934	

The result is an interpreted description containing 15 features of the ordered nanoparticles arrangement for each SEM image in the dataset:

- (O_1) General consistency of orientations;
- (O_2) Partial consistency of orientations;
- (O_3) The fraction of reliable orientations;
- (L_1) Number of lines constructed of the SUP;
- (L_2) Smoothness of the SUP-lines;
- (L_3) Rectilinearity of the SUP-lines;
- (L_4) The fraction of connected nanoparticles of the SUP-lines;
- (B_1^E, B_1^{PD}) Smoothness of the MSF-lines;
- (B_2^E, B_2^{PD}) Rectilinearity of the MSF-lines;
- (B_3^E, B_3^{PD}) The fraction of connected nanoparticles of the MSF-lines;
- (B_4^E, B_4^{PD}) The mean normalized length of the MSF-lines.

However, in previous work [10] it was shown that the use of only three features (O_2, O_3, L_4) makes it possible to obtain an acceptable classification quality. Therefore, it is worth reducing the resulting set of features. Note that features based on the minimum spanning forest calculated using various metrics can be combined by averaging the corresponding values (such features are denoted by \bar{B}). Tables 4 show the main quality indicators on 5-fold cross-validation for the linear SVM classifier and the reduced feature space.

Tab. 4. Linear SVM classifier quality indicators for $O_2, O_3, L_4, \bar{B}_1 - \bar{B}_4$ features

Target class	Acc	Precision	Recall	F	P_4
Ordered	0.984	0.990	0.990	0.989	0.973
Disordered		0.968	0.968	0.969	

From the results shown in Table 4, we can see that after reducing the set of features, the classification quality has improved significantly. Thus, it is sufficient to use only 7 interpreted features to analyze the ordering SEM images.

5.3. Model implementation and training

When calculating the proposed features of SEM images, the following parameter values were set.

For O -features (subsection 4.1):

- The proportionality coefficient for early stopping in local groups formation: $k = 3$;
- The weight coefficient for estimating the local nanoparticles density in a SEM image: $w_d = 1.5$;
- The maximum number of nanoparticles in a local group: $s = 8$.

For L -features (subsection 4.2):

- The reliability threshold for computing the partial consistency of orientations: $q_{min} = 0.85$;
 - Proportionality coefficient to adjust the degree of influence of individual parts of the proposed metric of prevailing directions: $C = 0.025$;
 - Weight coefficient of the angular coaxiality in the metric of prevailing directions to ensure line smoothness: $w_{coax} = 1.75$;
 - Minimum line length in nanoparticles: $L_{min} = 12$.
- For B -features (subsection 4.3):
- The maximum distance at which a nanoparticle can be connected in a graph: $dist_{max} = 20$;
 - The number of longest lines: $N_l = 10$;
 - Minimum line length in nanoparticles to account for bound nanoparticles: $N_n = 5$.

The size of a polyline local fragment is used to estimate the lines smoothness for L - and B - features: $f_{size} = 7$.

The number of lines detected in an image depends not only on its nature, but also on the settings of the search algorithm. These settings can range from searching for all possible lines, as shown in Fig. 5, to searching for none at all. We have chosen the optimal settings for these parameters in order to ensure a significant number of long lines in ordered images like the image on Fig. 1c, and to minimize the number of short lines in disordered images like the image on Fig. 1d.

The proposed algorithms were implemented in Python.

Based on the corresponding methods from the scikit-learn [25] package, the following steps were implemented: calculations of the prevailing directions of nanoparticles local groups and of the corresponding reliabilities (principal component analysis - decomposition.PCA); training of a linear SVM classifier (svm.SVC: the core is linear, the regularization parameter is 10); evaluation of the classifier quality (cross-validation - model_selection.StratifiedKFold: the number of folds is 5).

Detection of nanoparticles was performed based on a parallel algorithm proposed in our previous work [26, 27].

The following algorithms have an author's implementation: an algorithm for the formation of local groups of nanoparticles; an algorithm for constructing smooth lines based on a modification of the shortest unclosed path method; an algorithm for constructing a minimum spanning forest.

Depending on the number of nanoparticles in the original SEM image, the operating time (excluding the detection stage) of the proposed implementation varies from a couple of seconds for ~1000 nanoparticles to several dozens of minutes for ~20,000 nanoparticles. In the dataset

under study, the most common number of nanoparticles in the SEM image corresponds to ~5000, which is processed in a few minutes. The indicated time costs correspond to calculations on a personal computer with the following characteristics: processor – Intel® Core™ i7-9700k (3.6/4.9 GHz); RAM – 16 Gb (DDR4, 3866 MHz); SSD: 256 Gb, operating system – Windows 10 x64. Parallel computing technologies were not used in this experiment.

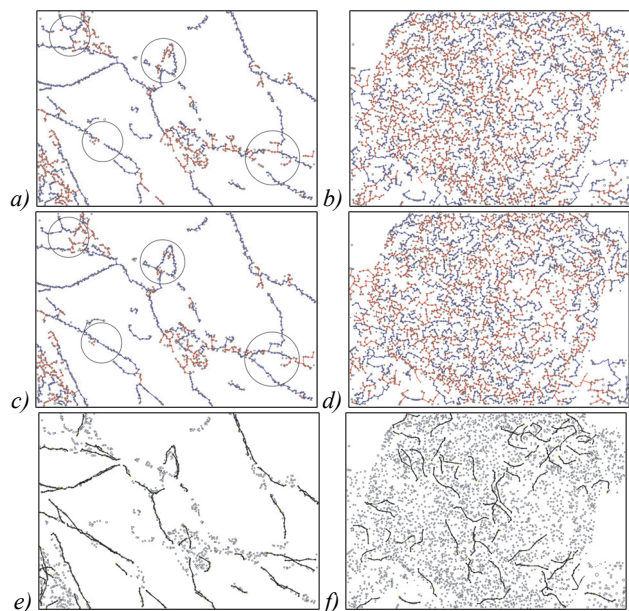


Fig. 5. The lines found based on the minimum spanning forest method for two typical SEM images using: (a), (b) – the Euclidean metric; (c), (d) – the metrics of Prevailing Directions. Blue lines are those whose length is greater than or equal to 7 nanoparticles. (e), (f) – the lines found based on the shortest unclosed path method (length of the lines is greater than or equal to 7 nanoparticles)

Conclusions

This paper presents the development of a previously proposed approach for determining the ordered and disordered arrangement of palladium nanoparticles in order to determine the distribution of defects on the surface of carbon materials. The approach includes calculating various simple characteristics from three main groups: the orientation of nanoparticles, the construction of broken lines using the shortest open path, and the minimum spanning forest. High classification quality values have been achieved, significantly exceeding the performance of more complex neural network models that were previously reported. Moreover, it has been demonstrated that in order to achieve high classification accuracy, a small number of order features is sufficient.

It is worth noting that the proposed approach, based on the analysis of explainable data, allows us to clearly interpret the classification results based on formalized features, which is not possible for a neural network that is represented as a "black box". This is significant because the proposed method can form the basis for a more general measure of orderliness – the degree of orderliness.

References

- [1] Titirici M-M, White RJ, Brun N, Budarin VL, Su DS, del Monte F, Clark JH, MacLachlan MJ. Sustainable carbon materials. *Chem Soc Rev* 2015; 44(1): 250-290. DOI: 10.1039/C4CS00232F.
- [2] Takakura A, Beppu K, Nishihara T, Fukui A, Kozeki T, Namazu T, Miyauchi Y, Itami K. Strength of carbon nanotubes depends on their chemical structures. *Nat Commun* 2019; 10: 3040. DOI: 10.1038/s41467-019-10959-7.
- [3] Morishita K, Takarada T. Scanning electron microscope observation of the purification behaviour of carbon nanotubes. *J Mater Sci* 1999; 34: 1169-1174. DOI: 10.1023/A:1004544503055.
- [4] Achaw O-W. A study of the porosity of activated carbons using the scanning electron microscope. In *Book: Kazmiruk V, ed. Scanning electron microscopy. Ch 24. InTech; 2012. Source: <https://www.intechopen.com/chapters/30949>. DOI: 10.5772/36337.*
- [5] Pentsak EO, Kashin AS, Polynski MV, Kvashnina KO, Glatzel P, Ananikov VP. Spatial imaging of carbon reactivity centers in Pd/C catalytic systems. *Chem Sci* 2015; 6: 3302-3313. DOI: 10.1039/C5SC00802F.
- [6] Pokrajac L, Abbas A, Chrzanowski W, Dias GM, Eggleton BJ, Maguire S, Maine E, Malloy T, Nathwani J, Nazar L, Sips A, Sone J, van den Berg A, Weiss PS, Mitra S. Nanotechnology for a sustainable future: Addressing global challenges with the international network4sustainable nanotechnology. *ACS Nano* 2021; 15(12): 18608-18623. DOI: 10.1021/acsnano.1c10919.
- [7] Zhang P, Guo Z, Ullah S, Melagraki G, Afantitis A, Lynch I. Nanotechnology and artificial intelligence to enable sustainable and precision agriculture. *Nat Plants* 2021; 7(7): 864-876. DOI: 10.1038/s41477-021-00946-6.
- [8] Jenewein KJ, Torresi L, Haghmoradi N, Kormanyos A, Friederich P, Cherevko S. Navigating the unknown with AI: multiobjective Bayesian optimization of non-noble acidic OER catalysts. *J Mater Chem A* 2024; 12: 3072-3083. DOI: 10.1039/D3TA06651G.
- [9] Boiko DA, Pentsak EO, Cherepanova VA, Gordeev EG, Ananikov VP. Deep neural network analysis of nanoparticle ordering to identify defects in layered carbon materials. *Chem Sci* 2021; 12(21): 7428-7441. DOI: 10.1039/D0SC05696K.
- [10] Kurbakov MY, Sulimova VV, Kopylov AV, Seredin OS, Boiko DA, Pentsak EO, Cherepanova VA, Ananikov VP. Determining the orderliness of carbon materials with nanoparticle imaging and explainable machine learning. *Nanoscale* 2024; 16(16): 13663-13676. DOI: 10.1039/d4nr00952e.
- [11] Boiko DA, Sulimova VV, Kurbakov MY, Kopylov AV, Seredin OS, Cherepanova VA, Pentsak EO, Ananikov VP. Automated recognition of nanoparticles in electron microscopy images of nanoscale palladium catalysts. *Nanomaterials* 2022; 12(21): 3914. DOI: 10.3390/nano12213914.
- [12] Prim RC. Shortest connection networks and some generalizations. *Bell Syst Tech J* 1957; 36(6): 1389-1401. DOI: 10.1002/j.1538-7305.1957.tb01515.x.
- [13] Gorban A, Kégl B, Wunch D, Zinovyev A, eds. *Principal manifolds for data visualization and dimension reduction*. Berlin, Heidelberg: Springer-Verlag; 2008. ISBN: 978-3-540-73749-0.
- [14] Hu M-K. Visual pattern recognition by moment invariants. *IRE Trans Inf Theory* 1962; 8(2): 179-187. DOI: 10.1109/TIT.1962.1057692.

- [15] Surkov EE, Seredin OS, Kopylov AV. Locally optimal solutions in the shortest unclosed path search problem. Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBREIT) 2023: 221-224. DOI: 10.1109/USBREIT58508.2023.10158834.
- [16] Nesetril J, Milkova E, Nesetrilova H. Otakar Boruvka on minimum spanning tree problem translation of both the 1926 papers, comments, history. Discrete Math 2001; 233(1-3): 3-36. DOI: 10.1016/S0012-365X(00)00224-7.
- [17] Kruskal JB. On the shortest spanning subtree of a graph and the traveling salesman problem. Proc American Mathematical Society 1956; 7: 48-50. DOI: 10.1090/S0002-9939-1956-0078686-7.
- [18] Boiko DA, Pentsak EO, Cherepanova VA, Ananikov VP. Electron microscopy dataset for the recognition of nanoscale ordering effects and location of nanoparticles. Sci Data 2020; 7(1): 101. DOI: 10.1038/s41597-020-0439-1.
- [19] Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948; 27(4): 623-656. DOI: 10.1002/j.1538-7305.1948.tb00917.x.
- [20] Cover TM, Thomas JA. Elements of Information Theory. Hoboken, New Jersey: Wiley; 1991. ISBN: 978-0-471-24195-9.
- [21] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. 14th Int Joint Conf on Artificial Intelligence 1995; 2: 1137-1143.
- [22] Esfahani MSh, Dougherty ER. Effect of separate sampling on classification accuracy. Bioinformatics 2014; 30(2): 242-250. DOI: 10.1093/bioinformatics/btt662.
- [23] Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Int J Mach Learn Technol 2011; 2(1): 37-63.
- [24] Seredin OS, Kopylov AV, Harmonic Averaging in Classifier Quality Assessment. Pattern Recognition and Image Analysis 2024; 34(4): 1160-1171.
- [25] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011; 12: 2825-2830.
- [26] Kurbakov MY, Sulimova VV. High-performance two-level parallel computing scheme for nanoparticles detection in SEM images. Int Arch Photogramm Remote Sens Spatial Inf Sci XLVIII-2/W3-2023: 145-150. DOI: 10.5194/isprs-archives-XLVIII-2-W3-2023-145-2023.
- [27] Kurbakov MYu, Sulimova VV, Seredin OS, Kopylov AV. A program for detecting nanoparticles in images from an electron microscope based on the exponential approximation method [In Russian]. Certificate of state registration of the computer program No. 2023688122 of December 20, 2023.

Author's information

Mikhail Yurievich Kurbakov completed Bachelor's (2019) and Master's (2021) degrees in Computer Science and Engineering from Tula State University (TSU). Currently, he is a postgraduate student at TSU with a degree in Mathematical and Software Support of Computing Systems, Complexes and Computer Networks and works as a junior researcher at the Laboratory of Cognitive Technologies and Simulating Systems (LCTSS), TSU. He is the author of more than 20 scientific papers and a number of application software. E-mail: muwsik@mail.ru

Valentina Vyacheslavovna Sulimova holds a Ph.D. in Theoretical Foundations of Computer Science from the Computing Center of the Russian Academy of Sciences, Moscow, Russia, in 2009. Currently, she is an Associate Professor at the Institute of Applied Mathematics and Computer Science, TSU, and is a Senior Researcher at the LCTSS, TSU. Her research interests include data mining, pattern recognition, machine learning, signal and image analysis, supercomputing technologies. She is a member of the International Association for Pattern Recognition (IAPR), a member of the reviewer board for SN Computer Science Journal and an expert of the Russian Science Foundation. Prof. Sulimova has authored more than 80 scientific papers in refereed journals and conference proceedings on machine learning, pattern recognition and supercomputing technologies. E-mail: vsulimova@yandex.ru

Oleg Sergeevich Seredin holds a Ph.D. in Theoretical Foundations of Computer Science from the Computing Center of the Russian Academy of Sciences, Moscow, Russia, in 2001. Currently, he is an Associate Professor at the Institute of Applied Mathematics and Computer Science, TSU, and is a Leading Researcher at the LCTSS, TSU. His research interests include data mining, pattern recognition, machine learning, signal and image analysis, visual representation of multidimensional data, and statistical methods for decision making. Prof. Seredin has served on the program committee of many conferences (CloudCom, AIST, GraphiCon, VISAPP, PSBB, PRIB, MaDaIn). He is also a member of the reviewer board for several journals, such as Sensors, Computer Optics, SN Computer Science Journal, IEEE Signal Processing Letters, Applied Science. He has worked as a visiting researcher at Rutgers University and National Taipei University of Technology. Prof. Seredin has authored more than 100 scientific papers in refereed journals and conference proceedings on machine learning, pattern recognition, and computer vision. He is a member of the IAPR. E-mail: oseredin@yandex.ru

Andrei Valerievich Kopylov received the Ph.D. degree from the Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russia, in 1997. Currently, he is an Associate Professor with the Institute of Applied Mathematics and Computer Science, TSU. Since 2022, he is also a Leading Researcher in the LCTSS, TSU. He worked as visiting researcher at the Dorodnicyn Computing Centre of Russian Academy of Sciences and National Taipei University of Technology. His scientific interests are signal and image analysis, data mining, machine learning. He is a member

of program committee at several conferences (CloudCom, PSBB, SoICT, AIST, GraphiCon, VISAPP, ICPR), reviewer of scientific journals Sensing and Imaging, Computer Optics, Machine Learning and Data Analysis, IEEE Signal Processing Letters, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Access, etc. He has published more than 100 scientific papers in refereed journals, handbooks, and conference proceedings in the areas of machine learning, pattern recognition and computer vision. Prof. Kopylov is a member of the IAPR. E-mail: and.kopylov@gmail.com

*Code of State Categories Scientific and Technical Information (in Russian – GRNTI): 28.23.15.
Received May 27, 2024. The final version – July 04, 2024.*
