

New Maximum Similarity Method for Object Identification in Photon Counting Imaging

V. E. Antsiperov ^a

Kotelnikov Institute of Radioengineering and Electronics of RAS, Mokhovaya 11-7, Moscow, Russian Federation

Keywords: Machine Learning, Image Recognition, Photon Counting Detectors, Poisson Point Process Intensity, Precedent-based Identification, Naive Bayes Model, Gaussian Mixture Model, K-means Clustering.

Abstract: The paper discusses a new approach to recognition / identification of the test objects according to their intensity shape in the images registered by photon counting detectors. The main problem analyzed within the framework of the proposed approach is related to the identification decision (inference) based on a registered set of discrete photocounts (~ photons) regarding the similarity of the shape of the object's intensity in the image to the shape of previously observed objects (precedents). It is shown that when the intensity shape is approximated by a mixture of Gaussian components within the framework of this approach, a recurrent identification algorithm can be synthesized, similar to the well-known K-means clustering algorithm in the machine (statistical) learning.


1 INTRODUCTION

Biomedical imaging includes several methods and techniques for representing by images (usually digital) various organs and their sections, hidden as a rule from direct (visual) observation. The demand for imaging tools in modern medicine is growing at a rapid pace, and this is noted both in the field of analysis of clinical cases and in the field of diagnostics. The need for modern diagnostic methods in medicine has led to the expansion of developments in various areas of biomedical imaging. Magnetic resonance imaging (MRI), computed (X-ray) tomography (CT), positron emission tomography (PET), single-photon emission computed tomography (SPECT), etc. should be noted among the main areas that have won reliable positions today (Darby, 2012).

The existence of various imaging methods is usually associated with the sensitivity of the corresponding techniques to a certain type of tissue. For example, MRI images are sensitive to soft tissue, while X-ray images are more sensitive to hard and bony structures. However, despite the difference in the underlying physical principles, a common feature of all methods is the low level of the used radiation.

This leads to the fact that the so-called photon-counting detectors (PCDs) are widely used as the sensors of the visualization (Leng, 2019). At the same time, while the photon-counting mode is natural for PET and SPECT, until recently, the mode of photons accumulation (energy-integrating detectors – EIDs) was mainly used for CT. However, the recent developments of energy-sensitive PCD open up new possibilities for obtaining X-ray images at low photon fluxes, involving the registration of images in the PCD mode (from ~ 10 photons per pixel to, in the future, 1:1) (Willeminck, 2018).

This energy-sensitive PCD technology has the potential to revolutionize clinical CT by providing a higher contrast-to-noise ratio, improving spatial resolution, and opening the possibilities for spectral (colour) imaging. In this regard, the report outlines a new image processing method that can be chosen as the basis for a modern approach to PCD-imaging problems. The proposed method is defined as the method of *maximum similarity* and represents some adaptation of the R. Fisher's maximum likelihood method (ML) to machine learning problems. Within the framework of the proposed approach, the recognition, or more precisely, the identification of objects on PCD images is performed in accordance with the shape of their intensity.

^a  <https://orcid.org/0000-0002-6770-1317>

2 MAXIMUM SIMILARITY METHOD

The starting points of the maximum similarity method formally are as follows. It is assumed that for the objects under consideration there are sets of (random) observations – counts of the form $X = \{\vec{x}_1, \dots, \vec{x}_n\}$, where $\vec{x}_i \in \mathbb{R}^2$. In this case, the number n of observations in set X can be arbitrary. The physical mechanism of their registration (observation) is not specified here. For the maximum similarity method presented in this work, only the statistical description of observations is essential.

From the statistical point of view, it is assumed that each observation \vec{x}_i is random and the process of its registration is described by some parametric probability distribution with density $\rho(\vec{x} | \vec{\theta})$, $\vec{\theta} \in \Theta \subset \mathbb{R}^p$ (parametric model). It is assumed, following the Bayesian point of view, that the parameters $\vec{\theta}$ are also random variables with a certain prior distribution density $\mathcal{P}(\vec{\theta})$, the exact form of which, however, is not essential. Both assumptions allow us to utilize the joint distribution density $\rho(\vec{x}; \vec{\theta}) = \rho(\vec{x} | \vec{\theta})\mathcal{P}(\vec{\theta})$ of observations \vec{x} and parameters $\vec{\theta}$.

The presented parametric model belongs to the class of the so-called generative models (Jebara, 2004), which imply a correspondence of certain values of the parameters $\vec{\theta}_1, \vec{\theta}_2, \dots, \vec{\theta}_k, \dots \in \Theta$ to the objects observed (in short, these values can be thought of as object class labels (Wasserman, 2004)). However, the problem with generative models is that the exact values of the parameters corresponding to the objects are unknown, only their statistical estimates, based on the sets of objects observed data $X_1, X_2, \dots, X_k, \dots$ are available.

In particular, the maximum likelihood (ML) estimates $\vec{\theta}_k^{(ML)}$ (Efron, 1982), determined from the R. Fisher's ML equations, can be used as such estimates:

$$\vec{\theta}_k^{(ML)} = \underset{\vec{\theta} \in \Theta}{\operatorname{argmax}} \rho(X_k | \vec{\theta}). \quad (1)$$

As follows from (1), to find maximum likelihood estimates, it is necessary for each n , as well as for $n = 1$, to determine its parametric n -model of the set of observations X – the corresponding probability distribution density $\rho(\vec{x}_1, \dots, \vec{x}_n | \vec{\theta})$ (these models for different n should be consistent in accordance with the Kolmogorov theorem, see (Billingsley, 1986)). However, it is possible to significantly simplify the problem of determining n -models by

assuming the conditional (for a given $\vec{\theta}$) independence of individual observations of the set:

$$\rho(X | \vec{\theta}) = \prod_{i=1}^n \rho(\vec{x}_i | \vec{\theta}). \quad (2)$$

Note that assumption (2) is actively used in machine learning problems, for example, within the framework of the naive Bayesian method (Barber, 2012), which is one of the ten most popular modern algorithms (Wu, 2007).

The problem of identification, correlation of some observable, test object with one of the previously registered training objects (hereinafter called precedents) is formalized now in the context of the presented parametric model as the problem of maximizing some measure of similarity of the observed data from the object $X = \{\vec{x}_1, \dots, \vec{x}_n\}$ with the data sets $X_1, X_2, \dots, X_k, \dots$, obtained earlier for the precedents. Since no additional knowledge about the object and precedents except for observed data $X, X_1, X_2, \dots, X_k, \dots$ is assumed (including the values of the characteristic parameters $\vec{\theta}_1, \vec{\theta}_2, \dots, \vec{\theta}_k, \dots \in \Theta$, characteristics of $\mathcal{P}(\vec{\theta})$, etc.), it is highly desirable that the corresponding similarity measure $\mu(X, X_k)$ could be expressed in terms of this and only this data.

A natural quantitative characteristic of the consistency of data X and an arbitrary set of observations X_k is the probability density of their joint distribution $p(X, X_k)$, under the assumption that both sets correspond to the same unknown object.

Taking into account the basic assumption about the considered model (2) (as well as the sets X and X_k conditional independence), density $p(X, X_k)$ can be written in the following form:

$$\begin{aligned} p(X, X_k) &= \int p(X, X_k | \vec{\theta}) \mathcal{P}(\vec{\theta}) d\vec{\theta} = \\ &= \int \rho(X | \vec{\theta}) \rho(X_k | \vec{\theta}) \mathcal{P}(\vec{\theta}) d\vec{\theta} = \\ &= \int \prod_1^n \rho(\vec{x}_i | \vec{\theta}) \prod_1^{n_k} \rho(\vec{x}_j | \vec{\theta}) \mathcal{P}(\vec{\theta}) d\vec{\theta} = \\ &= \int \rho(X \cup X_k | \vec{\theta}) \mathcal{P}(\vec{\theta}) d\vec{\theta} \end{aligned} \quad (3)$$

where $\{\vec{x}_j\}$ is a set of n_k observations X_k , $X \cup X_k$ is a set of $n + n_k$ observations obtained by uniting X and X_k . In other words, $p(X, X_k)$ (3) is an unconditional distribution density of type (2) model for united set of observations $\{\vec{x}_j\} \cup \{\vec{x}_i\}$. Note that the data X and X_k , being conditionally independent, in the general case turn out to be unconditionally dependent due to a correlation $X_k \sim \vec{\theta} \sim X$.

Considering the above interpretation, it is clear that $p(X, X_k)$ in some sense reflects the degree of consistency between X and X_k . Namely, if all objects observed data $X_1, X_2, \dots, X_k, \dots$, would be of the same size $n_1 = n_2 = \dots = n_k = m$, then all the sets $\{X \cup X_k\}$ would be random samples within the same

parametric $(n + m)$ -model $\varrho(X | \vec{\theta})$ (2). In this case, the degree of consistency $X \sim X_k$ would be determined by the probability of their joint sample $X \cup X_k$, i.e. $p(X, X_k)$, indeed, could be considered as a similarity measure. The problem, however, is that, due to the arbitrary value of n_k , the sets $\{X \cup X_k\}$ belong to different models and, therefore, the comparison of $p(X, X_k)$ values for different X_k should be corrected for this circumstance.

In the proposed maximum similarity method, the corresponding correction is specified by normalizing the values $p(X, X_k)$ (3) to the probabilities $p(X_k)$:

$$\mu(X, X_k) = \frac{p(X, X_k)}{p(X_k)} = p(X | X_k). \quad (4)$$

i.e. the similarity measure $\mu(X, X_k)$ is chosen as the ratio of the probability of the $X \cup X_k$ sample to the probability of the only X_k sample. In this case, the maximum similarity method consists in choosing the precedent for which the expansion of the sample X_k by the X leads to the greatest probability ratio.

As follows from (4), the chosen probability ratio formally coincides with the conditional probability of the set X with the given observations X_k . The latter leads to an alternative interpretation of the proposed method: the observed object is identified with the precedent, observations X_k of which lead to the maximum conditional probability of a set of observations X for the object tested.

Using the selected measure of similarity of the observed data and precedents $\mu(X, X_k)$ (4), the maximum similarity method can be formalized as a solution to the following maximum similarity (MS) equation:

$$k_{MS} = \underset{k}{\operatorname{argmax}} \mu(X, X_k) = \underset{k}{\operatorname{argmax}} p(X | X_k). \quad (5)$$

To further substantiate the proposed method, namely, the choice of the similarity measure in the form (4), it is convenient to turn to the asymptotic case of large samples of precedents $n_k \gg 1$. Note that, in addition to questions of convenience, this case quite adequately reflects the specifics of the (machine) learning process organization.

In the general (not necessarily asymptotic) case the conditional probability $p(X | X_k)$ in terms of the parametric model (3) can be written as:

$$p(X | X_k) = \frac{\int \varrho(X | \vec{\theta}) \varrho(X_k | \vec{\theta}) \mathcal{P}(\vec{\theta}) d\vec{\theta}}{\int \varrho(X_k | \vec{\theta}) \mathcal{P}(\vec{\theta}) d\vec{\theta}}. \quad (6)$$

In the case $n_k \gg 1$ for $\varrho(X_k | \vec{\theta})$ (2) holds the well-known asymptotic approximation in neighbourhood of $\vec{\theta}_k^{(ML)}$ (1) (Wasserman, 2004):

$$\varrho(X_k | \vec{\theta}) \cong \varrho(X_k | \vec{\theta}_k^{(ML)}) \times \exp \left[-\frac{n_k}{2} (\vec{\theta} - \vec{\theta}_k^{(ML)})^T I(\vec{\theta}_k^{(ML)}) (\vec{\theta} - \vec{\theta}_k^{(ML)}) \right]. \quad (7)$$

where T is the transposition operation, $I(\vec{\theta}_k^{(ML)})$ is the Fisher's information matrix for the distribution $\varrho(\vec{x} | \vec{\theta})$ – one of the most important characteristics of the adopted parametric model:

$$I_{ij}(\vec{\theta}) = - \int \varrho(\vec{x} | \vec{\theta}) [\partial^2 \ln \varrho(\vec{x} | \vec{\theta}) / \partial \theta_i \partial \theta_j] d\vec{x}. \quad (8)$$

Considering the sharpness of the peak of asymptotic (7) in the vicinity of $\vec{\theta}_k^{(ML)}$, we can approximate the numerator in (6) by:

$$\int \varrho(X | \vec{\theta}) \varrho(X_k | \vec{\theta}) \mathcal{P}(\vec{\theta}) d\vec{\theta} \approx \varrho(X | \vec{\theta}_k^{(ML)}) \int \varrho(X_k | \vec{\theta}) \mathcal{P}(\vec{\theta}) d\vec{\theta}. \quad (9)$$

As a result, $p(X | X_k)$ (6) is simplified to $\varrho(X | \vec{\theta}_k^{(ML)})$, and the similarity measure $\mu(X, X_k)$ (4) takes the following simple form:

$$\mu(X, X_k) \cong \varrho(X | \vec{\theta}_k^{(ML)}). \quad (10)$$

In other words, in the case of large sets of precedent observations $X_1, X_2, \dots, X_k, \dots, n_k \gg 1$ (the number of observations n of a set X of an identified object does not have to be large), the naive Bayesian distribution density $\varrho(X | \vec{\theta})$ (2) for the values of the parameter $\vec{\theta} = \vec{\theta}_k^{(ML)}$ can be used as a similarity measure $\mu(X, X_k)$. The latter means that for arbitrary precedents (arbitrary n_k), the similarity measures $\mu(X, X_k)$ (4) are determined within the same n -model, so the comparison of their values is quite justified. Note that not only ML estimates but also many other, called consistent estimates, give us an asymptotic of the form (7). In this context, $\vec{\theta}_k^{(ML)}$ can be understood as any of consistent estimates, and not necessarily a solution to problem (1).

With a pragmatic point of view, expression (10) for the similarity measure $\mu(X, X_k)$ seems even more attractive than (4). The corresponding formulation of the maximum likelihood method, which is the asymptotic limit of the general formulation (5), takes on a form like the maximum likelihood method (1):

$$k_{MS} = \underset{k}{\operatorname{argmax}} \varrho(X | \vec{\theta}_k^{(ML)}) = \underset{\vec{\theta} \in \{\vec{\theta}_k^{(ML)}\}}{\operatorname{argmax}} \varrho(X | \vec{\theta}). \quad (11)$$

with the only exception that maximization is performed not over all $\vec{\theta} \in \theta$, but only over a finite set of estimates $\vec{\theta}_1^{(ML)}, \vec{\theta}_2^{(ML)}, \dots, \vec{\theta}_k^{(ML)}, \dots \in \theta$. Note that, from a practical point of view, in this case, there is also no need to store in full the sets of observations of

precedents $X_1, X_2, \dots, X_k, \dots$, it is enough to save only the parameter estimates (statistics) $\{\vec{\theta}_k^{(ML)}\}$, obtained from observations.

3 MAXIMUM SIMILARITY METHOD IN THE CASE OF GAUSSIAN MIXTURES

In previous works (Antsiperov, 2019a), (Antsiperov, 2019b) basing on the physical (semi-classical) mechanisms of radiation registration by matter, it was shown that the most adequate statistics of photocounts in the PCD image data $X = \{\vec{x}_1, \dots, \vec{x}_n\}$, is given by the Poisson point processes (PPP) model (Streit, 2010). It was also shown that for the problems of object identification only by the form of intensity, regardless of the total brightness, one can restrict oneself to the densities of conditional distributions of the probabilities of the coordinates of the counts $\{\vec{x}_i\}$ factored into the product, provided that their total number of registered counts n is given:

$$\begin{aligned} \varrho(\vec{x}_1, \dots, \vec{x}_n | n, I(\vec{x})) &= \prod_{i=1}^n \rho(\vec{x}_i | I(\vec{x})), \\ \rho(\vec{x} | I(\vec{x})) &= I(\vec{x}) / \int_{\Omega} I(\vec{x}) d\vec{x}, \end{aligned} \quad (12)$$

where $I(\vec{x})$ is the intensity of coming from the object and recorded on the PCDs surface Ω radiation.

Let the intensity $I(\vec{x})$ be approximated by a parametric intensity model $I(\vec{x} | \vec{\theta}_0)$, $\vec{\theta}_0 \in \theta \subset \mathbb{R}^p$, which we define as a sum, a mixture of K overlapping components, frames $\{F_j(\vec{x} | \vec{\mu}_j, \vec{\theta}_0)\}$, $j = 1, \dots, K$, located at the nodes $\{\vec{\mu}_j\}$ of some imaginary regular lattice, covering Ω :

$$I(\vec{x} | \vec{\theta}_0) = \sum_{j=1}^K F_j(\vec{x} | \vec{\mu}_j, \vec{\theta}_0). \quad (13)$$

The components $F_j(\vec{x} | \vec{\mu}_j, \vec{\theta}_0)$ in (13) are assumed to be copies of some basic frame $F(\vec{x} | \vec{\eta}) \geq 0$, $\vec{\eta} \in \mathbb{R}^p$, $p = P/K$, shifted to the nodes of the lattice $\{\vec{\mu}_j\}$, $p = P/K$, specified up to the values of the parameters $\vec{\eta}$. The region $\Delta \subset \mathbb{R}^2$, in which $F(\vec{x} | \vec{\eta}) \neq 0$ will be called the carrier of the base frame or the base carrier. This carrier Δ is assumed to be symmetric in the sense that it contains the coordinates origin $\vec{x} = \vec{0}$ and, together with each $\vec{x} \in \Delta$, contains $-\vec{x}$. The simplest example of such a carrier is some regular polygon, for example, a square, placed in coordinates origin.

Let us denote by $\vec{\eta}_j$ the values of the parameters of the frame / component in the mixture (13), located

at the node j ($\vec{\eta}_j$ can be considered as related to this node j as $\vec{\mu}_j$). Taking into account the assumptions made, model (1) is refined in the form:

$$I(\vec{x} | \vec{\theta}_0) = \sum_{j=1}^K F(\vec{x} - \vec{\mu}_j | \vec{\eta}_j). \quad (14)$$

where the complete set of parameters $\vec{\theta}_0$ is now represented by the set $\{\vec{\eta}_1, \dots, \vec{\eta}_K\}$ and it is considered that the j -th component of the mixture depends only on a part of the parameters – on $\vec{\eta}_j$.

In model (14), the carriers of the components of the mixture $\{\Delta_j\}$ – copies of the base carrier Δ shifted by $\{\vec{\mu}_j\}$, are assumed to be partially overlapping. This is provided by the requirement $D > d$, where D is the characteristic size of Δ , and d is the lattice spacing. Assuming the last requirement is fulfilled, we obtain that the set of carriers $\{\Delta_j\}$ completely covers the area Ω of the image. The simplest example of such a covering Ω is the square carriers $\{\Delta_j\}$, whose centers are located at the nodes $\{\vec{\mu}_j\}$ of a rectangular lattice.

When Ω is completely covered by the carriers $\{\Delta_j\}$, each point $\vec{x} \in \Omega$ belongs to at least one carrier. Therefore, the set of nodes whose carriers contain \vec{x} is not empty. We denote the set of indices of these nodes by $\delta_{\vec{x}} = \{j\}$ and call it the \vec{x} lattice environment. Due to the symmetry of the base carrier Δ , the nodes in $\delta_{\vec{x}}$ will be those contained in the region obtained by displacement to the point \vec{x} of the base carrier Δ . Lattice environments $\delta_{\vec{x}}$ will be used intensely in the construction of identification procedure.

The next step in refining model (14) is related to the choice of a dependence of the base frame $F(\vec{x} | \vec{\eta})$ on the parameters $\vec{\eta} = (\eta_0, \eta_1, \dots, \eta_{p-1})^T$. The simplest form of such dependence could be a linear combination with the coefficients $\vec{\eta}$ of some finite functional basis $\{\varphi_0(\vec{x}), \varphi_1(\vec{x}), \dots, \varphi_{p-1}(\vec{x})\}$, for example, in the image of the construction of frames (Gröchenig, 2001). However, in this case, to ensure the positivity of the base frame $F(\vec{x} | \vec{\eta}) \geq 0$, significant restrictions would be required both on the basis $\{\varphi_q(\vec{x})\}$ and on the parameters $\vec{\eta}$. Therefore, it seems more appropriate to expand on the functional basis of not the frame itself, but its logarithm:

$$F(\vec{x} | \vec{\eta}) = I_0 \exp\{\sum_{q=0}^{p-1} \eta_q \varphi_q(\vec{x})\}, \vec{x} \in \Delta, \quad (15)$$

where the multiplier I_0 is introduced to ensure the correct frame dimension. Bearing in mind the subsequent normalization of the intensity, it is convenient to take as multiplier I_0 the average radiation intensity on the Ω :

$$I_o = \frac{1}{\Sigma_\Omega} \iint_\Omega I_o(\vec{x}) d\vec{x}, \quad (16)$$

where Σ_Ω is the surface area of Ω .

It is natural to require that the parametric family $F(\vec{x} | \vec{\eta})$ (15) contains at least all the constants $I(\vec{x}) \equiv I > 0$. For this, one of the basic functions, for example $\varphi_0(\vec{x})$, should be admitted as a constant: $\varphi_0(\vec{x}) \equiv 1$. The corresponding parameter η_0 will set the normalization of the components by means of the factor $\exp\{\eta_0\}$ in (15). In view of the subsequent transition to the normalized version of the intensity, it is convenient to introduce instead of the parameter η_0 another parameter $w = w(\vec{\eta})$, which is a function of $\vec{\eta}$ and has the meaning of the energy fraction per frame (15) for the given $\vec{\eta}$ of a total (falling per unit time on Ω) energy $W_O = I_o \Sigma_\Omega$:

$$\begin{aligned} w(\vec{\eta}) &= \frac{1}{W_O} \iint_\Delta F(\vec{x} | \vec{\eta}) d\vec{x} = \\ &= \exp\{\eta_0\} \frac{1}{\Sigma_\Omega} \iint_\Delta \exp\left\{\sum_{q=1}^{p-1} \eta_q \varphi_q(\vec{x})\right\} d\vec{x} = \quad (17) \\ &= \exp\{\eta_0\} \exp\{A(\eta_1, \dots, \eta_{p-1})\} \end{aligned}$$

where an auxiliary (cumulant-generating function) function related to the basis $\{\varphi_q(\vec{x})\}$ is introduced:

$$\begin{aligned} A(\eta_1, \dots, \eta_{p-1}) &= \\ &= \ln \left\{ \frac{1}{\Sigma_\Omega} \iint_\Delta \exp\left\{\sum_{q=1}^{p-1} \eta_q \varphi_q(\vec{x})\right\} d\vec{x} \right\}. \quad (18) \end{aligned}$$

Replacing the parameters $\vec{\eta} \rightarrow (w, \eta_1, \dots, \eta_{p-1})$ in accordance with (17), we arrive at the following representation of the base frame (15):

$$\begin{aligned} F(\vec{x} | w, \vec{\eta}) &= \\ &= I_o w \exp\{\vec{\eta} \cdot \vec{\varphi}(\vec{x}) - A(\vec{\eta})\} \Pi_\Delta(\vec{x}). \quad (19) \end{aligned}$$

where the shortened notations $\vec{\eta} = (\eta_1, \dots, \eta_{p-1})^T$ and $\vec{\varphi}(\vec{x}) = (\varphi_1(\vec{x}), \dots, \varphi_{p-1}(\vec{x}))^T$ are introduced, symbol of dot product “ \cdot ” is used and the characteristic function $\Pi_\Delta(\vec{x})$ of the base carrier Δ is introduced to automatically take into account the constraint $\vec{x} \in \Delta$.

Considering the chosen structure of the base frame $F(\vec{x} | w, \vec{\eta})$ (19), the parametric model of intensity $I(\vec{x} | \vec{\theta}_O)$ (14) takes the following final form:

$$\begin{aligned} I(\vec{x} | \vec{\theta}_O) &= I_o \sum_{j=1}^K w_j \times \\ &\times \exp\{\vec{\eta}_j \vec{\varphi}(\vec{x} - \vec{\mu}_j) - A(\vec{\eta}_j)\} \times \quad (20) \\ &\times \Pi_\Delta(\vec{x} - \vec{\mu}_j) \end{aligned}$$

where $\{w_j\}$ and $\{\vec{\eta}_j\}$ – are weights and sets of normal parameters of j -th components, $\vec{\varphi}(\vec{x})$ is a functional basis of the parametric model common for all components, I_o and $A(\vec{\eta}_j)$ are defined, respectively, by expressions (16) and (18). Note that, if we formally calculate the integral over Ω of the right-hand side of (20), taking into account (18), the will get the value $I_o \Sigma_\Omega \sum_{j=1}^K w_j$, which implies the property that the weights $\{w_j\}$ are normalized to unity, which coincides with their interpretation as the distribution of the total radiation energy $W_O = I_o \Sigma_\Omega$ over individual components.

Using the worked out parametric model of the recorded radiation intensity (20), we write down the probability distribution density (12) of an individual observation count from the sample $X_n = \{\vec{x}_1, \dots, \vec{x}_n\}$, representing the of the object O image, also in the parametric form:

$$\begin{aligned} \rho(\vec{x} | I(\vec{x})) &= I(\vec{x} | \vec{\theta}_O) / \int_\Omega I(\vec{x} | \vec{\theta}_O) d\vec{x} = \\ &= \frac{1}{\Sigma_\Omega \sum_{j \in \delta_{\vec{x}}} w_j} \exp\{\vec{\eta}_j \vec{\varphi}(\vec{x} - \vec{\mu}_j) - A(\vec{\eta}_j)\} \times \quad (21) \\ &\times \Pi_\Delta(\vec{x} - \vec{\mu}_j) \end{aligned}$$

where it is considered that for a given count \vec{x} the formal summation over all components j from 1 to K in (20) is actually reduced to summation over the nonzero components at the point \vec{x} – over the lattice environment $\delta_{\vec{x}}$.

Even though all subsequent conclusions can be made in the most general case of mixtures of distributions of an exponential family, for simplicity of presentation and to avoid cumbersome formulas, we restrict ourselves to the case of a simple model of Gaussian mixtures (GMM) (Murphy, 2012).

Namely, we assume that the functional basis $\vec{\varphi}(\vec{x})$ contains only two linear basis functions $\varphi_1(\vec{x}) = x_1/D$ and $\varphi_2(\vec{x}) = x_2/D$ (so the number of $\vec{\eta}$ components is also two: $\vec{\eta} = (\eta_1, \eta_2)$). In addition, we approximate the characteristic function $\Pi_\Delta(\vec{x})$ by a Gaussian bell-shaped distribution $\exp\{-\vec{x}^2/2D^2\}$. Replacing the integration in (18) with $\Pi_\Delta(\vec{x})$ by integrating with Gaussian approximation, we find that an auxiliary function $A(\vec{\eta})$ has a quite simple form:

$$A(\vec{\eta}) = \frac{1}{2} \vec{\eta}^2 + \ln \left[\frac{2\pi D^2}{\Sigma_\Omega} \right]. \quad (22)$$

Substituting $A(\vec{\eta})$ from (22) into the expression for $\rho(\vec{x} | I(\vec{x}))$ (21) we finally obtain a representation of the classical model of a Gaussian mixture (GMM) mixtures (Murphy, 2012) for the probability distributions of individual counts from the sample X_n :

$$(\vec{x}|I(\vec{x})) = \frac{1}{2\pi D^2} \sum_{j \in \delta_{\vec{x}}} w_j \exp \left\{ -\frac{(\vec{x} - \vec{\mu}_j - D\vec{\eta}_j)}{2D^2} \right\} \quad (23)$$

In the chosen model (23), corresponding n -model of the observations $X_n = \{\vec{x}_1, \dots, \vec{x}_n\}$ (2) after opening all n brackets in the product takes the form:

$$\begin{aligned} \varrho(X_n | \vec{\theta}_0) &= \prod_{i=1}^n \left[\sum_{j \in \delta_{\vec{x}_i}} \frac{w_j}{2\pi D^2} \exp \left\{ -\frac{(\vec{x}_i - \vec{\mu}_j - D\vec{\eta}_j)}{2D^2} \right\} \right] \\ &= \sum_{j_1 \in \delta_{\vec{x}_1}} \dots \sum_{j_n \in \delta_{\vec{x}_n}} \prod_{i=1}^n \left[\frac{w_{j_i}}{2\pi D^2} \times \right. \\ &\quad \left. \times \exp \left\{ -\frac{(\vec{x}_i - \vec{\mu}_{j_i} - D\vec{\eta}_{j_i})}{2D^2} \right\} \right] \end{aligned} \quad (24)$$

where the indices $j_i \in \delta_{\vec{x}_i}$ associate the samples \vec{x}_i with nodes j of the lattice environment $\delta_{\vec{x}_i}$, the summation is performed over all possible n -tuples $\{j_1, \dots, j_n\} \in \delta_{\vec{x}_1} \times \dots \times \delta_{\vec{x}_n} = \delta_{X_n}$.

Note that each of the n -tuples $\{j_1, \dots, j_n\} \in \delta_{X_n}$ defines a partition of the sample $X_n = \{\vec{x}_1, \dots, \vec{x}_n\}$ into K disjoint subsets of counts $\{\pi_j\}$ associated with nodes j : $\pi_j = \{\vec{x}_i | j_i = j\}$, $X_n = \bigcup_{j=1}^K \pi_j$. If in the product terms of the sum (24) we combine the exponents and give similar terms with respect to the parameters w_j and $\vec{\eta}_j$, then the density $\varrho(X_n | \vec{\theta}_0)$ can be rewritten as a sum over all partitions:

$$\begin{aligned} \varrho(X_n | \vec{\theta}_0) &= \sum_{\{j_1, \dots, j_n\} \in \delta_{X_n}} \left[\exp \left\{ \sum_{j=1}^K \bar{p}_j \ln w_j \right\} \right]^n \\ &\times \left[\exp \left\{ -\frac{\sum_{j=1}^K \bar{p}_j (\vec{\Phi}_{\pi_j} - D\vec{\eta}_j)^2}{2D^2} \right\} \right]^n \times \\ &\times \left[\frac{\exp \left\{ -\frac{1}{2D^2} \sum_{j=1}^K \bar{p}_j C_{\pi_j} \right\}}{2\pi D^2} \right]^n \end{aligned} \quad (25)$$

where $\bar{p}_j = n_j/n$ are the partition weights of nodes, $n_j = |\pi_j|$ is the number of samples \vec{x}_i in the subset π_j , corresponding node j , $\sum_{j=1}^K n_j = n$, $\sum_{j=1}^K \bar{p}_j = 1$ and $\{\vec{\Phi}_{\pi_j}\}$, $\{C_{\pi_j}\}$ are defined by the partition $\{\pi_j\}$ as follows:

$$\begin{aligned} \vec{\Phi}_{\pi_j} &= \frac{\sum_{\vec{x}_i \in \pi_j} (\vec{x}_i - \vec{\mu}_j)}{n_j}, \\ C_{\pi_j} &= \frac{\sum_{\vec{x}_i \in \pi_j} (\vec{x}_i - \vec{\mu}_j - \vec{\Phi}_{\pi_j})^2}{n_j}. \end{aligned} \quad (26)$$

Obviously, when the total number of counts n is large, the calculation of all $|\delta_{X_n}| \sim k^n$ terms in (24) / (25), where $k > 1$ is the average size of the lattice environment $\delta_{\vec{x}}$, $\vec{x} \in \Omega$, becomes a difficult problem.

For example, in the case when the size of the base carrier D noticeably exceeds the lattice spacing d , $k \approx (D/d)^2 \gg 1$, the number of terms becomes extremely large, and the problem becomes an *EXP*-complete problem (Du, 2014).

Therefore, the problem of approximating $\varrho(X_n | \vec{\theta}_0)$ (24) / (25), with something simpler is urgent. One of the possibilities here is, obviously, the approximation of this big sum with its only term. A remarkable fact is that if the chosen term is maximal compared to others in some neighbourhood of some $\vec{\theta}^{(*)}$, then as the number of counts n grows, the term becomes the asymptotic approximation of the distribution $\varrho(X_n | \vec{\theta})$ (25) in $\vec{\theta}^{(*)}$. Indeed, if we denote the terms in (25), corresponding to the partitions as $[s_{\{\pi_j\}}(\vec{\theta})]^n$, and by $\{\pi_j^{(*)}\}$ we denote the partition corresponding to the maximal in $\vec{\theta}^{(*)}$ term, then (25) can be rewritten as:

$$\begin{aligned} \varrho(X_n | \vec{\theta}) / [s_{\{\pi_j^{(*)}\}}(\vec{\theta})]^n &= \\ = 1 + \sum_{\{\pi_j\} \neq \{\pi_j^{(*)}\}} [s_{\{\pi_j\}}(\vec{\theta}) / s_{\{\pi_j^{(*)}\}}(\vec{\theta})]^n \end{aligned} \quad (27)$$

Since by assumption in some neighbourhood of $\vec{\theta}^{(*)}$ all ratios in the right-hand sum (27) are less than unity, so if we denote the average of degree n of all $\sim k^n$ ratios by ε_n , then we also get $\varepsilon_n < 1$. So far as the sum in (27) can be estimated from above by $[k\varepsilon_n]^n$, for $k\varepsilon_n < 1$, as the number of counts n in the sample X_n grows, this sum will tend to zero and, therefore, the asymptotic $\varrho(X_n | \vec{\theta}) / [s_{\{\pi_j^{(*)}\}}(\vec{\theta})]^n \rightarrow 1$ will take place. The latter means that the maximal term for $n \gg 1$ acquires a dominant value in the vicinity of $\vec{\theta}^{(*)}$.

Considering the above reasoning, the problem of approximation $\varrho(X_n | \vec{\theta})$ is thus reduced to the question of how to find the partition $\{\pi_j^{(*)}\}$ for which the corresponding term in (25) will be maximal in the neighbourhood of its most probable, optimal parameter $\vec{\theta}^{(*)} = \{\{w_j\}, \{\vec{\eta}_j\}\}$. It is clear, that the problem of finding such partition $\{\pi_j^{(*)}\}$ / optimal $\vec{\theta}^{(*)}$ as direct comparison of the maxima of all $\sim k^n$ terms of sum (25) is not suitable, since it is also an *EXP*-complete problem (Du, 2014). Fortunately, there are quite effective procedures for solving optimization problems like (25). Below we propose one of such procedures – recurrent segmentation / partition of sample $X_n = \{\vec{x}_1, \dots, \vec{x}_n\}$, analogous to the well-known K -means segmentation method (Barber, 2012), (Wu, 2007). This procedure consists of

sequential (recurrent) iterations of two main steps – step P for refining the target partition $\{\pi_j^{(v)}\}$ and step M – for calculating the corresponding $\{\pi_j^{(v)}\}$ optimal parameters $\vec{\theta}^{(v)}$.

Refinement at step P of the partition $\{\pi_j^{(v)}\}$ with the parameters $\vec{\theta}^{(v-1)}$ found at the previous iteration is carried out as a solution for each sample $\vec{x}_i \in X_n$ of the following maximization problem:

$$j_i = \arg \max_{j \in \delta_{\vec{x}_i}} \left[w_j^{(v-1)} \exp \left\{ -\frac{(\vec{x}_i - \vec{\mu}_j - D\vec{\eta}_j)}{2D^2} \right\} \right], \quad (28)$$

where the set of nodes j tested for maximum is limited by the \vec{x}_i -th lattice environment $\delta_{\vec{x}_i}$.

The solution j_i of each of the problems (28) selects in the sum (25) a subset of those terms for which the i -th factor is greater than that of the others. After the indices are found for all counts $\{\vec{x}_1, \dots, \vec{x}_n\}$, the n -tuples $\{j_1, \dots, j_n\} \in \delta_{X_n}$ will define some refined partition $\{\pi_j^{(v)}\}$ the corresponding term of which for the values parameters $\vec{\theta}^{(v-1)}$ will be maximum in the sum (24) / (25).

With the partition $\{\pi_j^{(v)}\}$ found at step P, finding at step M the corresponding optimal parameters $\vec{\theta}^{(v)}$ is the quite simple problem. First, the weights $\{w_j\}$ are included in each term of (25) only in the expression for the cross-entropy $-\sum_{j=1}^K \bar{p}_j \ln w_j$ in first factor and give it a maximum at $w_j = \bar{p}_j$ for all j . Second, the maximum of each term with respect to $\vec{\eta}_j$ is obvious from the form of second factor in (25) and is achieved at $\vec{\eta}_j = \vec{\Phi}_{\pi_j} / D$. Collecting these facts into a system, we find that at step M the following calculations should be performed:

$$\begin{aligned} w_j^{(v)} &= \bar{p}_j^{(v)} = |\pi_j^{(v)}| / n, \\ \vec{\eta}_j^{(v)} &= \vec{\Phi}_{\pi_j^{(v)}} / D = \frac{1}{|\pi_j^{(v)}| D} \sum_{\vec{x}_i \in \pi_j^{(v)}} (\vec{x}_i - \vec{\mu}_j) \end{aligned} \quad (29)$$

The question of stopping the recurrent procedure (28)–(29) is solved in the same way as in the case of K -means segmentation, i.e. is determined by the criterion for stabilization of partitions: $\{\pi_j^{(v)}\} = \{\pi_j^{(v-1)}\} = \{\pi_j^{(*)}\}$. Note that the convergence of the partitions $\{\pi_j^{(v)}\} \rightarrow \{\pi_j^{(*)}\}$ is accompanied by the convergence of the parameters $\vec{\theta}^{(v)} \rightarrow \vec{\theta}^{(*)}$, but it does not follow from this that in the limit we obtain the maximum likelihood estimate $\vec{\theta}^{(ML)}$ (1).

As a result, calculating $\vec{\theta}^{(*)}$ (or, equivalently $\{\pi_j^{(*)}\}$) and approximating the density $\varrho(X_n | \vec{\theta}_0)$ (25) with the corresponding term, we obtain, according to

(10), the following approximation for the similarity measure:

$$\begin{aligned} \mu(X_n, X_k) &= Q(\vec{\theta}^{(*)}) [\exp\{-D_{KL}(\{w_j^*, \{w_j^k\})\})\}]^n \\ &\times \left[\exp \left\{ -\frac{1}{2} \sum_{j=1}^K w_j^{(*)} (\vec{\eta}_j^{(*)} - \vec{\eta}_j^k)^2 \right\} \right]^n \end{aligned} \quad (30)$$

where $\vec{\theta}_k^{(ML)} = \{\{w_j^k\}, \{\vec{\eta}_j^k\}\}$ are parameters of the precedent X_k and $\{w_j^{(*)}\}$ and $\{\vec{\eta}_j^{(*)}\}$ are the optimal parameters of the tested sample X_n . In (30) the model parameters \bar{p}_j and $\vec{\Phi}_{\pi_j}$ (26), dependent on the partition, were replaced in accordance with (29) with the optimal parameters calculated during the PM procedure. For the convenience, the cross-entropy $-\sum_{j=1}^K \bar{p}_j \ln w_j$ in (25) is replaced by the sum of the Kullback–Leibler divergence $D_{KL}(\{w_j^*, \{w_j\})$ (Murphy, 2012) and the Shannon's entropy $H(\{w_j^*\})$:

$$\begin{aligned} D_{KL}(\{w_j^*, \{w_j\}) &= -\sum_{j=1}^K w_j^* \ln w_j / w_j^*, \\ H(\{w_j^*\}) &= -\sum_{j=1}^K w_j^* \ln w_j^*. \end{aligned} \quad (31)$$

Finally, all independent of the precedent X_k factors in $\varrho(X_n | \vec{\theta}_0)$ are combined in (30) into one common $Q(\vec{\theta}^{(*)})$:

$$\begin{aligned} Q(\vec{\theta}^{(*)}) &= \left[\frac{\exp\{H(\{w_j^*\})\}}{2\pi D^2} \right]^n \times \\ &\times \left[\exp \left\{ -\frac{1}{2D^2} \sum_{j=1}^K w_j^{(*)} C_{\pi_j^{(*)}} \right\} \right]^n \end{aligned} \quad (32)$$

Theoretically, the similarity measure (30) explicitly sets the required expression for finding the most similar precedent to the tested object (11). However, the computational characteristics of the method can be noticeably improved if any equivalent similarity measure $\tilde{\mu}(X_n, X_k)$ – monotonic function of $\mu(X_n, X_k)$ will be used. Namely, discarding the first factor $Q(\vec{\theta}^{(*)})$ in (30), which does not depend at all on the parameters $\vec{\theta}_k^{(ML)}$, and taking the natural logarithm of the second factor, up to a factor n we obtain the following similarity measure:

$$\begin{aligned} \tilde{\mu}(X_n, X_k) &= -D_{KL}(\{w_j^*, \{w_j^k\}) - \\ &- \frac{1}{2} D_{ME}(\{\vec{\eta}_j^{(*)}\}, \{\vec{\eta}_j^k\}) \end{aligned} \quad (33)$$

where we have introduced the notation D_{ME} for the average over the distribution of $\{w_j^*\}$ Euclidean distance between the corresponding parameters of $\{\vec{\eta}_j^{(*)}\}$ and $\{\vec{\eta}_j^k\}$:

$$D_{ME}(\{\tilde{\eta}_j^{(*)}\}, \{\tilde{\eta}_j^k\}) = \sum_{j=1}^K w_j^{(*)} (\tilde{\eta}_j^{(*)} - \tilde{\eta}_j^k)^2 \quad (34)$$

Thus, in the case of modeling the intensities of the registered radiation by a simple model of Gaussian mixture (24) / (25), the maximum similarity method is reduced to the choice of the minimum sum of divergence between the optimal parameters of the tested object $\tilde{\theta}^{(*)}$ and the maximum likelihood parameters of the precedents $\tilde{\theta}_k^{(ML)}$. Denoting the sum of divergences (31), (33) by $\bar{\mu}(X, X_k) = -\tilde{\mu}(X, X_k)$, we can reformulate the maximum similarity method (11) for this case as follows:

$$\begin{aligned} k_{MS} &= \arg \min_k \bar{\mu}(X_n, X_k) = \\ &= \arg \min_k \left[D_{KL}(\{w_j^*\}, \{w_j^k\}) + \frac{1}{2} D_{ME}(\{\tilde{\eta}_j^{(*)}\}, \{\tilde{\eta}_j^k\}) \right] \end{aligned} \quad (35)$$

which in the above entry is literally the criterion of minimum difference.

4 CONCLUSIONS

The considered case of approximating the shape of the intensity of objects by mixtures of Gaussian components and the results of the corresponding numerical simulation showed the adequacy of the method of maximum similarity to the problems of analyzing PCD images. Even in low-quality images (~ 1000 samples), the algorithm corresponding to the proposed method gives the correct identification of objects. Note that the implementation of the algorithm, like the method of K -means segmentation, is very efficient computationally – for mixtures with ~ 1000 components in the common computation environment, processing of images with a size of 500×500 pixels by PM algorithm (28) – (29) takes ~ 1 sec on a standard PC and it is already clear that these characteristics can be improved if desired.

As for the maximum similarity method itself, the simplicity of its interpretation, to which the section 2 is devoted, and the straightforwardness of its algorithmic implementation, what is the main content of the section 3, makes it attractive both in theoretical and practical terms, especially in the context of modern, oriented to machine learning approaches. In a sense, for machine learning problems, the proposed method is an adaptation of the R. Fisher's maximum likelihood method widely used in traditional statistics (Efron, 1982). The fruitful use of the latter, as is known, has led to a huge number of important statistical results. In this regard, it is hoped that the proposed maximum similarity method will also be

useful in solving a wide range of modern problems of statistical (machine) learning.

ACKNOWLEDGEMENTS

The author is grateful to the Russian Foundation for Basic Research (RFBR), grant N 18-07-01295 A for the financial support of the work.

REFERENCES

- Darby, M.J., Barron, D.A., Hyland, R.E., 2012. Chapter 1. Techniques. In *Oxford Handbook of Medical Imaging*. Oxford University Press. Oxford.
- Leng, S., et. al. 2019. Photon-counting Detector CT: System Design and Clinical Applications of an Emerging Technology. In *RadioGraphics*, 39(3): 729-743.
- Willemink, M.J., et. al., 2018. Photon-counting CT: Technical Principles and Clinical Prospects. Un *Radiology*, 289(2): 293-312.
- Gonzalez, R.C., Woods, R.E., 2007. *Digital Image Processing*, Prentice Hall. 3rd edition.
- Saleh, B., 1978. *Photoelectron Statistics*. Springer Verlag. Berlin.
- Antsiperov, V., 2019a. Machine Learning Approach to the Synthesis of Identification Procedures for Modern Photon-Counting Sensors. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods – ICPRAM*. SCITEPRESS.
- Jebara, T., 2004. *Machine Learning: Discriminative and Generative*. Kluwer. Dordrecht.
- Wasserman, L., 2004. *All of Statistics: A Concise Course in Statistical Inference*. Springer. New York.
- Efron, B., 1982. Maximum likelihood and decision theory. In *Ann. Statist.*, 10: 340–356.
- Billingsley, P., 1986. *Probability and Measure*, Wiley. New York. 2nd edition.
- Barber, D., 2012. *Bayesian Reasoning and Machine Learning*, Cambridge Univ. Press. Cambridge.
- Wu, X., et. al., 2007. Top 10 algorithms in data mining. In *Knowl. Inf. Syst.* 14 (1): 1–37.
- Antsiperov, V.E., 2019b. Target identification for photon-counting image sensors, inspired by mechanisms of human visual perception. In *Journal of Physics: Conference Series*, 1368: 032020.
- Streit, R.L., 2010. *Poisson Point Processes. Imaging, Tracking and Sensing*. Springer. New York.
- Gröchenig, K., 2001. *Foundations of Time-Frequency Analysis*. Birkhäuser. Boston.
- Murphy, K.P., 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press. Cambridge.
- Du, D., Ko, K., 2014. *Theory of Computational Complexity*, 2nd ed. John Wiley & Sons.