Kolpakov A.V.
Bauman Moscow State Technical University

Design of Algorithm for Pigmented Nevi Border Form Classification

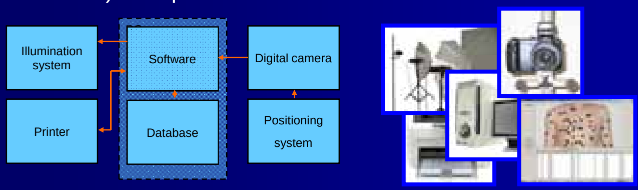
Urgency of the Problem

• Incidence of malignant melanoma has increased dramatically over the past 20 years.
• There is a necessity of diagnostic methods improvement and automation of pigmented nevi image analysis for early diagnosis of premalignant transformations of the skin.

Fig. 1. Surface spreading melanoma

Target Setting

- The task of algorithm design for pigmented skin nevi classification is solving in the frames of solving global task of skin mapping hardware and software complex “Rhodonite” (SMHSC “Rhodonite”) development.

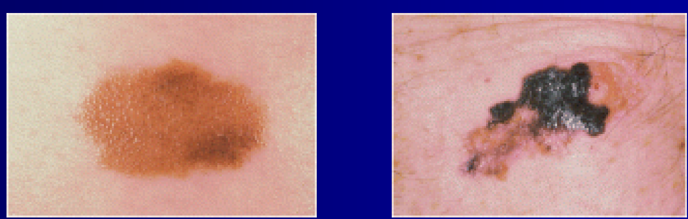


a b

Fig. 2. (a) block diagram of skin mapping hardware and software complex “Rhodonite”; (b) components of “Rhodonite”.

Target Setting

- In the first place algorithm for nevi border form classification is designing.



a b

Fig. 3. (a) Benign nevus (regular border); (b) Melanoma (irregular border)

Description

1. Selection of Border Description Method

Signature is the method of object border description with univariate function that may be constructed in different ways [4]. The reasons for using this method of border description are:

- possibility to reduce border description to univariate function that easier to describe;
- possibility to construct signature that will be invariant to parallel transfer, rotation and scaling.

The border will be submitted like a consistency of the pairs of coordinates on the plane surface *xy*:

$$s(k) = (x(k),y(k)). \tag{1}$$

Then each pair of coordinates we able to consider as a complex number:

$$s(k) = x(k)+iy(k) \tag{2}$$

Function of complex variation *s(k)* will be border signature.

Description

- In the capacity of border descriptors we will use elements of complex number consistency a(u), that gained from the consistency s(k) by mean of Fourier transform:

$$a(u) = \sum_{k=0}^{N-1} s(k) e^{\frac{i2\pi uk}{N}} \tag{3}$$

- Complex numbers a(u) named Fourier-descriptors [4]. Quadrates of Fourier-descriptors' magnitudes contain information about form of the border and may be used as a base for border form classification.

Description

2. Interpolation of an Input Data

For ensuring independent from border perimeter length of number array, containing consistency of Fourier-descriptors, the interpolation of an input data is necessary.

Length of the array will determine according to next considerations:

- Quantity of elements in the array N must be multiple of *2k*, where *k* is the whole number. This condition gives possibility to optimize calculation of Fourier transform.
- Number array must enclose the longest consistency of Fourier-descriptors that was met in sample.

According to mentioned criterions, array length 256 elements were empirically determined. Then linear interpolation of an input data was applied: the quantity of countings was changed according to scale factor *K= N/Nin* (*Nin* – length of an input consistency) for ensuring needed length of consistency.

$$Q_k = (\mu_{1k} - \mu_{2k}) \Big/ \left[(1/N_1)\sum_{j=1}^{N_1} \left| x_{1jk} - \mu_{1k} \right| + (1/N_2)\sum_{j=1}^{N_2} \left| x_{2jk} - \mu_{2k} \right| \right],$$

Description

3. Vector of Characteristics Forming

Getting of vector of characteristics from vector of an input data mathematically may be written as:

$$\boldsymbol{Y} = \boldsymbol{AX} \qquad (4)$$

Where ***X*** – vector of an input data with dimension *L*1*; ***Y*** – vector of characteristics with dimension *M*1* (ML); ***A*** – transform matrix *M*L*. As an algorithm for extraction of characteristics the algorithm of projection to one dimension (POD) was selected.

POD algorithm [5] allows selecting for space of characteristics such components, for which difference between average data vectors to overall dispersion inside the classes ratio is sufficient. Weight factor for *k* coordinate is defined as

$$Q_k = (\mu_{1k} - \mu_{2k}) \Big/ \left[(1/N_1)\sum_{j=1}^{N_1} \left| x_{1jk} - \mu_{1k} \right| + (1/N_2)\sum_{j=1}^{N_2} \left| x_{2jk} - \mu_{2k} \right| \right], \qquad (5)$$

where *μik* – *k* component of average data vector *μi* ; *Ni*- number of sample vectors, that belong to class *i*; *xik*- - *k* component of sample *j* of class *i*[5]. *Q* values are calculating for all *N* components in initial space.

Description

4. Selection of Classification Method

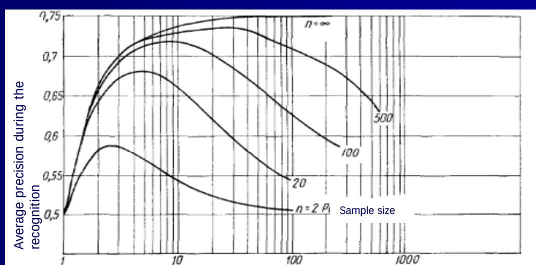
- There was statistical law experimental study conducted for each characteristic for selection of classification method.
- The study showed, that there is unavailable to plot segregating surface in selected space of characteristics that would have a simple form.
- The simplest rule of classification, applicable in such case is decisive rule of k nearest neighbors.
- We chose the metrical algorithm of classification with variable width of window that is k nearest neighbors method modification.

Description

5. Approval of the Algorithm

There were formed samples for each class of objects for the algorithm approval. Each sample was separated on test and learning parts. Sample size was selected according to considerations of prevention of "curse of dimensionality" occurrence.

Influence of the sample size on the validity of the classification is shown at the figure 4.



Dimensionality of space of characteristics

Fig. 4. Dependence of classification precision from dimensionality of space of characteristics

Results

- Designed classifier recognized correctly 41 of 60 objects from the test sample of objects with irregular border. Thereby probability of error during the recognition the object with irregular border (type I error) amount $P_I = 32\ \%$.
- Designed classifier recognized correctly 53 of 60 objects from the test sample of objects with regular border. Thereby probability of error during the recognition the object with regular border (type II error) amount $P_{II} = 11.7\ \%$.
- Acceptable in clinical practice values of type I and type II errors amount *20 %* [16]. Though designed algorithm provides high probability of type I error it is necessary to improve the algorithm. However, designed algorithm provides significant specificity:

$$Sp = 1 - P_{II} = 88{,}3\ \%$$

So it may be applicable to confirm the diagnosis because it provides seldom positive result in the absence of disease.