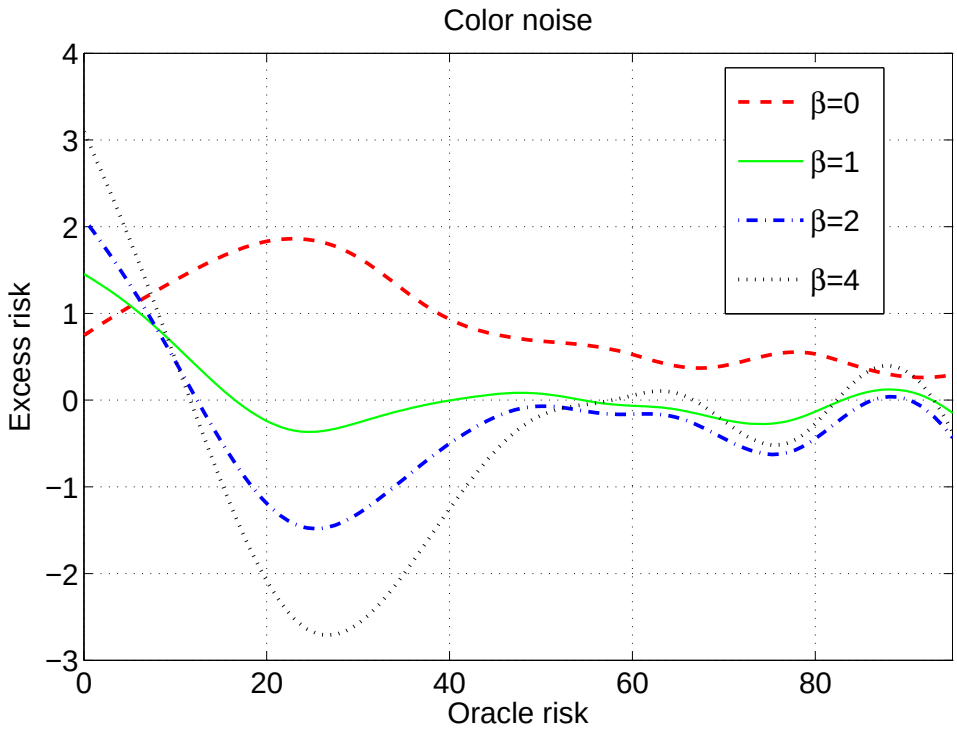
ISSN 2223-3792

Машинное обучение и анализ данных

2015 год **Том 1, номер 13**



Машинное обучение и анализ данных

Журнал «Машинное обучение и анализ данных» публикует новые теоретические и обзорные статьи с результатами научных исследований в области искусственного интеллекта, теоретических основ информатики и приложений. Цель журнала — развитие теории машинного обучения, интеллектуального анализа данных и методов проведения вычислительных экспериментов. Принимаются статьи на английском и русском языках.

Журнал включен в российский индекс научного цитирования РИНЦ. Информация о цитировании статей находится на сайте Российского индекса научного цитирования, ISSN 2223-3792, номер свидетельства о регистрации ЭЛ № ФС 77-55486. Журнал зарегистрирован в системе Crossref, doi http://dx.doi.org/10.21469/22233792.

- Новостной сайт http://jmlda.org/
- Электронная система подачи статей http://jmlda.org/papers/
- Правила подготовки статей http://jmlda.org/papers/doc/authors-guide.pdf

Тематика журнала:
— классификация, кластеризация, регрессионный анализ;
— алгебраический подход к проблеме синтеза корректных алгоритмов;
— многомерный статистический анализ;
— выбор моделей и сложность;
— глубокое обучение;
— статистическая теория обучения;
— методы прогнозирования временных рядов;
— методы обработки и распознавания сигналов;
— методы оптимизации в задачах машинного обучения и анализа данных;
— методы анализа больших данных;
— методы визуализации данных;
— обработка и распознавание речи и изображений;
— анализ и понимание текста;
— информационный поиск;
— прикладные задачи анализа данных.

Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Москва, 2015

Journal of Machine Learning and Data Analysis

The journal Machine Learning and Data Analysis publishes original research papers and reviews of the developments in the field of artificial intelligence, theoretical computer science and its applications. The journal aims to promote the theory of machine learning and data mining and methods of conducting computational experiments. Papers are accepted in English and Russian.

The journal is included in the Russian science citation index RSCI. Information about citation to articles can be found at the Russian science citation index website. ISSN 2223-3792. Mass media registration certificate ЭЛ № ФС 77-55486. The Crossref journal doi is `http://dx.doi.org/10.21469/22233792`.

- Journal news and archive http://jmlda.org/
- Open journal system for papers submission http://jmlda.org/papers/
- Style guide for authors http://jmlda.org/papers/doc/authors-guide.pdf

The scope of the journal:
— classification, clustering, regression analysis;
— multidimensional statistical analysis;
— Bayesian methods for regression and classification;
— model selection and complexity;
— deep learning;
— Statistical Learning Theory;
— time series forecasting techniques;
— methods of signal processing and speech recognition;
— optimization methods for solving machine learning and data mining problems;
— methods of big data analysis;
— data visualization techniques;
— methods of image processing and recognition;
— text analysis, text mining and information retrieval;
— applied data analysis problems.

Dorodnicyn Computing Centre FRC CSC RAS
Moscow Institute of Physics and Technology
Department of Control and Applied Mathematics
Division “Intelligent Systems”

Moscow, 2015

Содержание

Е. А. Крымова
Агрегация упорядоченных оценок в цветном шуме 1775
Е. В. Медведева, И. С. Трубин, Е. А. Устюжанина, А. В. Лалетин
Нелинейная многомерная фильтрация многокомпонентных изображений 1786
Н. Г. Федотов, А. А. Семов, А. В. Моисеев
Минимизация признакового пространства распознавания трехмерного изображения на основе стохастической геометрии и функционального анализа 1796
Л. А. Бекларян, Н. К. Хачатрян
Динамическая модель организации грузоперевозок 1815
Е. А. Новиков, М. А. Падалко
Использование Радон и Фурье преобразований растровых изображений для описания и отслеживания заданных объектов 1827
Е. П. Петров, Н. Л. Харина, Е. Д. Ржаникова
Комбинированная нелинейная фильтрация цифровых изображений большой разрядности 1844
Е. В. Дюкова, А. Г. Никифоров
Об эффективном распараллеливании алгоритмов для дискретных перечислительных задач 1853
И. В. Бахмутова, В. Д. Гусев, Л. А. Мирошниченко, Т. Н. Титкова
Параллельные тексты в задаче дешифровки древнерусских знаменных песнопений 1866
Д. А. Молчанов, Д. А. Кондрашкин, Д. П. Ветров
Машина релевантных тегов 1877

Contents

E. A. Krymova
Aggregation of ordered smoothers in colored noise 1775
E. V. Medvedeva, I. S. Trubin, E. A. Ustyuzhanina, and A. V. Laletin
Multidimensional nonlinear filtration of multicomponent images 1786
N. G. Fedotov, A. A. Syemov, and A. V. Moiseev
Feature space minimization of three-dimensional image recognition based on stochastic geometry and functional analysis 1796
L. A. Beklaryan and N. K. Khachatryan
Dynamic model of organization of cargo transportation 1815
E. A. Novikov and M. A. Padalko
The use of Radon and Fourier transformations of raster images for description and tracking of predefined objects 1827
E. P. Petrov, N. L. Kharina, and E. D. Rzhanikova
Combined nonlinear filtration of digital halftone high bitness images 1844
E. V. Djukova and A. G. Nikiforov
On efficient parallelizing of the algorithms for discrete enumeration problems 1853
I. V. Bakhmutova, V. D. Gusev, L. A. Miroshnichenko, and T. N. Titkova
Parallel texts in the problem of deciphering of ancient Russian chant 1866
D. A. Molchanov, D. A. Kondrashkin, and D. P. Vetrov
Relevance tagging machine 1877

Aggregation of ordered smoothers in colored noise*

E. A. Krymova
krymova@phystech.edu
Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Russia;
Institute for Information Transmission Problems, 19 Bolshoy Karetny per., build 1, Moscow, Russia

The paper is devoted to the problem of recovery of one-dimensional functions given a set of noisy observations. Suppose that in addition, one is given a fixed set of a finite number of function estimates. Based on this set of estimates, it is necessary to construct a new estimator, the risk of which would be close to the risk of the "best" estimate (so-called oracle) in a given set. The "best" estimator is a minimizer of the risk over the given set of function estimators. New oracle inequalities for aggregation of regression function estimates in assumption of heteroscedasic Gaussian noise, namely, correlated Gaussian noise with different variances at each design point, have been proved.

Keywords: *aggregation; exponential weighting; ordered smoothers; unbiased risk estimation*

DOI: 10.21469/22233792.1.13.01

Агрегация упорядоченных оценок в цветном шуме*

Е. А. Крымова
Московский физико-технический институт (ГУ)
Институт проблем передачи информации им. А. А. Харкевича РАН

Рассматривается задача восстановления функции регрессии по конечному числу наблюдений функции в гауссовским шуме, заданных в конечном числе детерминированных точек. Предположим, что помимо наблюдений функции исследователю заранее известен фиксированный набор из конечного числа оценок функции. На основе этого набора оценок требуется построить новую оценку, качество которой было бы сравнимо с наилучшей (в смысле среднеквадратичного риска) оценкой из заданного множества (с так называемым «оракулом»). В работе получены новые оракульные неравенства для экспоненциальной агрегации упорядоченных оценок функции регрессии в предположении гетероскедастичного шума, а именно: шум предполагается коррелированным (ковариационная матрица известна) и дисперсия его различна в каждой точке наблюдения.

Ключевые слова: *агрегация оценок; экспоненциальное взвешивание; упорядоченные оценки; несмещенное оценивание риска*

DOI: 10.21469/22233792.1.13.01

1 Introduction

The paper is devoted to estimation of noisy vector (sequence space model) given a set of linear estimators. The sequence space model plays significant role in nonparametric statistics. Many problems can be transformed to the sequence space model formulation with white (i. e., with noncorrelated identically distributed zero-mean noise) Gaussian noise or with colored (i. e.,

*This work is partially supported by RFBR research project 15-07-09121.

noncorrelated nonidentically distributed zero-mean) Gaussian noise. For example, very often, linear inverse problems are easily transformed into diagonal form with the help of singular value decomposition [1]. In this paper, the generalization of such models for the correlated colored Gaussian noise assumption is considered. Throughout the paper, it is assumed that one is given a special set of linear estimators, namely, ordered smoothers as various methods in statistics can be proved to have properties of ordered smoothers (for example, smoothing splines [2, 3], spectral regularization methods [1, 4], etc.). There exist various approaches to construct estimates given a set of estimators. One can use a model selection approach and select one estimator, for example, by a method of the unbiased risk estimation [5] which goes back to [6, 7].

Another approach is to use aggregation, namely, using a convex combination of given estimators. This approach was firstly developed by Nemirovsky [8] and independently by Catoni [9]. To tune the weights of the linear combination, authors performed the sample splitting. Later, this method was extended to several statistical models (see, e. g., [10–13]).

One can avoid sample splitting with the help of the exponential weighting. This method originates from the solution of functional aggregation problem by penalized empirical risk minimization [12]. It has been shown that for this method, one can yield rather good oracle inequalities for certain statistical models [14–16].

The goal is to prove new oracle inequalities for aggregation of ordered smoothers in assumption of heteroscedasic Gaussian noise, namely, correlated Gaussian noise with different variances at each design point.

2 Problem Statement

This paper deals with a sequence space model

$$Y_i = \theta_i + \xi_i, \quad i = 1, \dots, n, \tag{1}$$

where $(Y_1, \dots, Y_n)^\intercal$ is the vector of observation; and $(\xi_1, \dots, \xi_n)^\intercal$ is the zero-mean Gaussian vector with known $n \times n$ covariance matrix Σ. The goal is to estimate an unknown vector $\theta \in \mathbb{R}^n$ based on the data $Y = (Y_1, \dots, Y_n)^\intercal$.

Denote the diagonal elements of Σ by σ_i^2, $i = 1, \dots, n$. Let one impose the following conditions on the covariance matrix Σ .

1. The spectral norm is bounded from above:

$$\sigma_{\max}^2 = \sup_{x \in \mathbb{R}^n,\ \|x\|=1} x^\intercal \Sigma x < \infty\,.$$

2. The smallest eigenvalue is bounded from below:

$$\sigma_{\min}^2 = \inf_{x \in \mathbb{R}^n,\ \|x\|=1} x^\intercal \Sigma x > 0\,.$$

3. Assume also that

$$\sup_{x \in \mathbb{R}^n,\ \|x\|=1} x^\intercal \left[\Sigma \circ \Sigma\right] x < C_\circ^2$$

where $\circ$ is the Hadamard product and $C_\circ$ is the constant.

Let one denote the risk of an estimator $\hat\theta(Y) = (\hat\theta_1(Y), \dots, \hat\theta_n(Y))^\intercal$ by

$$R(\hat\theta, \theta) = \mathsf{E}_\theta \|\hat\theta(Y) - \theta\|^2. \tag{2}$$

Here, E_θ stands for the expectation with respect to the measure P_θ generated by the observations (1) where $\|\cdot\|$ denotes the norm in $\mathbb{R}^n$: $\|x\|^2 = \sum_{i=1}^n x_i^2$.

Throughout this paper, θ will be recovered with the help of linear estimates

$$\hat{\theta}_i^h(Y) = h_i Y_i,\ h \in \mathcal{H} \tag{3}$$

where $\mathcal{H}$ is the finite set of so-called *ordered smoothers*, which has the following definition.

Definition 1. *A set $\mathcal{H}$ is a set of ordered multipliers if*

- $h_i \in [0,1],\ i = 1,\dots,n$ *for all* $h \in \mathcal{H}$;
- $h_{i+1} \leqslant h_i,\ i = 1,\dots,n$, *for all* $h \in \mathcal{H}$; *and*
- *if for some integer* k *and some* $h, g \in \mathcal{H}$, $h_k < g_k$, *then* $h_i \leqslant g_i$ *for all* $i = 1,\dots,n$.

The last condition means that vectors in $\mathcal{H}$ are naturally ordered, since for any $h, g \in \mathcal{H}$, there are only two possibilities: $h_i \leqslant g_i$ or $h_i \geqslant g_i$ for all $i = 1,\dots,n$.

Substituting the linear model (3) into the risk definition (2), one obtains

$$R(\hat{\theta}^h, \theta) = \|(1-h)\cdot\theta\|^2 + \|\sigma\cdot h\|^2,$$

where $x \cdot y$ denotes the coordinate-wise product of vectors $x, y \in \mathbb{R}^n$, i. e., $z = x \cdot y$ means that $z_i = x_i y_i$, $i = 1,\dots,n$, and $\sigma = (\sigma_1,\dots,\sigma_n)^\intercal$. Since $R(\hat{\theta}^h, \theta)$ depends on $h \in \mathcal{H}$, one can minimize it over $h \in \mathcal{H}$. The minimal risk

$$r^{\mathcal{H}}(\theta) = \min_{h\in\mathcal{H}} R(\hat{\theta}^h, \theta)$$

is often called in the literature as the oracle risk [8, 9].

Naturally, it is not possible to use the estimate

$$\theta^*(Y) = h^* \cdot Y, \quad h^* = \arg\min_{h\in\mathcal{H}} R(\hat{\theta}^h, \theta)$$

because it depends on the unknown vector θ. But if one knew θ, it would be possible to point out the estimate with the least risk. That is why, the goal is to construct an estimator $\tilde{\theta}^{\mathcal{H}}(Y)$ based on the family of linear estimators $\hat{\theta}^h(Y),\ h \in \mathcal{H}$, which is close to the oracle risk. Formally, this means that the estimator $\tilde{\theta}^{\mathcal{H}}(Y)$ should satisfy the so-called oracle inequality

$$R(\tilde{\theta}^{\mathcal{H}}, \theta) \leqslant r^{\mathcal{H}}(\theta) + \tilde{\Delta}^{\mathcal{H}}(\theta)$$

which holds uniformly in $\theta \in \mathbb{R}^n$.

This inequality implies that the term $\tilde{\Delta}^{\mathcal{H}}$ is small with respect to the oracle risk uniformly in $\theta \in \mathbb{R}^n$. It is well known that in general, it is not possible to construct such an estimator [17]. But as it was shown in [17] for the set $\mathcal{H}$ of *ordered smoothers*, one can find an estimator which provides the following properties of the remainder term:

- $\tilde{\Delta}^{\mathcal{H}}(\theta) \leqslant \tilde{C} r^{\mathcal{H}}(\theta)$ for all $\theta \in \mathbb{R}^n$ where $\tilde{C} > 1$ is the constant; and
- $\tilde{\Delta}^{\mathcal{H}}(\theta) \ll r^{\mathcal{H}}(\theta)$ for all $\theta :\ r^{\mathcal{H}}(\theta) \gg \sigma^2$.

That is why, throughout this paper, it will be assumed that the set $\mathcal{H}$ contains solely ordered multipliers. Below, an example of ordered smoothers is given. Note that ordered smoothers are very common in statistics, e. g., smoothing splines [2, 3], spectral regularization methods [1, 4].

3 A Motivating Example

Consider the regression estimation problem in the case of colored noise. It is necessary to recover a one-dimensional function $f(x),\ x \in [0,1]$, given the noisy observations

$$Z_i = f(x_i) + \bar{\xi}(x_i), \quad i = 1,\dots,n, \tag{4}$$

where $x_i \in (0,1)$ and $\bar{\xi}_i(x)$ is the centered Gaussian random process with variance $\bar{\sigma}^2(x)$. Denote by $\bar{\Sigma}$ the covariance matrix of the vector $(\bar{\xi}(x_1), \dots, \bar{\xi}(x_n))^\intercal$.

Let one make use of the smoothing spline estimate, which is defined as follows:

$$\hat{f}_\alpha(x, Z) = \arg\min_f \left\{ \sum_{i=1}^n [Z_i - f(x_i)]^2 + \alpha \int_0^1 [f^{(m)}(x)]^2 \right\} \tag{5}$$

where $f^{(m)}(\cdot)$ denotes the derivative of order m and $\alpha > 0$ is the smoothing parameter which is usually chosen with the help of the Generalized Cross Validation (see, e. g., [18]).

To transform this model into the model (1), consider the Demmler–Reinsch basis [19] $\psi_k(x),\ x \in [0,1],\ k = 1, \dots, n$, which has double orthogonality property

$$\langle \psi_k, \psi_l \rangle_n = \delta_{kl}\,;$$

$$\int_0^1 \psi_k^{(m)}(x) \psi_l^{(m)}(x)\, dx = \delta_{kl} \lambda_k,\ k, l = 1, \dots, n,$$

where here and below $\langle u, v \rangle_n$ stands for the inner product

$$\langle u, v \rangle_n = \frac{1}{n} \sum_{i=1}^n u(x_i) v(x_i)$$

and λ_i are the eigenvalues of the basis.

It is assumed for definiteness that the eigenvalues λ_k are sorted in ascending order:

$$\lambda_1 \leqslant \dots \leqslant \lambda_n.$$

With this basis, one can represent the underlying function as follows:

$$f(x) = \sum_{k=1}^n \psi_k(x) \theta_k \tag{6}$$

and one gets from (4)

$$Y_k = \langle Z, \psi_k \rangle_n = \theta_k + \xi_k$$

where

$$\xi_k = \sum_{j=1}^n \bar{\xi}(x_k) \psi_k(x_j)\,. \tag{7}$$

Next, substituting (6) in (5), one arrives at

$$\hat{f}_\alpha(x, Z) = \arg\min_f \left\{ \sum_{k=1}^n (Y_k - \theta_k)^2 + \alpha \sum_{k=1}^n \lambda_k \theta_k^2 \right\}.$$

Therefore,

$$\hat{f}_\alpha(x, Y) = \sum_{k=1}^n \hat{\theta}_k \psi_k(x)$$

where

$$\hat{\theta}_k = \frac{Y_k}{1 + \alpha \lambda_k}.$$

Thus, one may conclude that the models (1)–(3) and (4)–(5) become equivalent with

$$h_k = h_k^\alpha = \frac{1}{1+\alpha\lambda_k}.$$

The vector $\xi = (\xi_1, \ldots, \xi_n)^\intercal$ is a Gaussian zero-mean vector with covariance matrix

$$\Sigma = \frac{1}{n^2}\Psi^\intercal\bar{\Sigma}\Psi$$

where matrix Ψ consists of the columns $(\psi_i(x_1), \ldots, \psi_i(x_n))^\intercal$, $i = 1, \ldots, n$.

From the orthogonality property of Demmler–Reinsch basis, it is easily seen that eigenvalues of matrix Σ are equal to $\sigma_i^2 = \bar{\sigma}^2(x_i)/n$. Thus, for fixed n, the matrix Σ has finite eigenvalues and the problem is equivalent to (1).

The most interesting case is when $\bar{\Sigma}$ is a diagonal matrix with diagonal elements $\bar{\sigma}^2(x_1), \ldots, \bar{\sigma}^2(x_n)$. It is known that in the case of equidistant design, Demmler–Reinsch basis has the following asymptotic as $n, k \to \infty$ [2]:

$$\psi_k(x) \approx \sqrt{\frac{2}{n}}\cos(\pi k x).$$

After a transformation of the regression estimation problem (4) with the help of Demmler–Reinsch basis, one obtains the following covariance of the noise (7):

$$\mathsf{E}\xi_k\xi_j \approx \frac{1}{n}\sum_{i=1}^{n}\sigma^2(x_i)\cos(\pi(k-j)p).$$

Thus, matrix Σ approximately equals to a correlation matrix of a stationary Gaussian sequence with variance $\sum_{i=1}^n \sigma^2(x_i)/n$ and the problem (4) becomes equivalent to the problem of estimation of an unknown vector in assumption of stationary noise.

In practice, one has to estimate the unknown covariance in (4). For the model with stationary noise, it is easy to estimate variance σ^2 given the data, for example, by

$$\bar{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^{n-1}[Z_i - Z_{i+1}]^2.$$

4 Exponential Weighing of Ordered Smoothers

In what follows, the exponential weighting estimate is used:

$$\bar{\theta}(Y) = \sum_{h\in\mathcal{H}} w^h(Y)\hat{\theta}^h(Y)$$

where

$$w^h(Y) = \pi^h \exp\left[-\frac{\bar{r}(Y,\hat{\theta}^h)}{2\beta\sigma^2_{\max}}\right] \Big/ \sum_{g\in\mathcal{H}} \pi^g \exp\left[-\frac{\bar{r}(Y,\hat{\theta}^g)}{2\beta\sigma^2_{\max}}\right].$$

Here, parameter $\beta > 0$ is fixed and $\bar{r}(Y,\hat{\theta}^h)$ is the unbiased risk estimate of $\hat{\theta}^h(Y)$ defined by

$$\bar{r}(Y,\hat{\theta}^h) \stackrel{\text{def}}{=} \|Y - \hat{\theta}^h(Y)\|^2 + 2\sum_{i=1}^{n} h_i\sigma_i^2 - \sum_{i=1}^{n}\sigma_i^2.$$

In order to cover $\mathcal{H}$ with small and large cardinalities, make use of the special prior weights defined as follows:

$$\pi^h \stackrel{\text{def}}{=} 1 - \exp\left\{-\frac{\sum_{i=1}^n \sigma_i^2(h_i^+ - h_i)}{\beta\sigma_{\max}^2}\right\}. \tag{8}$$

Here,

$$h^+ = \min\{g \in \mathcal{H} : g > h\}, \quad \pi^{h_{\max}} = 1$$

where $h^{\max}$ is the maximal multiplier in $\mathcal{H}$. Along with these weights, one needs also the following condition which can be proved to be true for smoothing splines and spectral regularization methods.

Condition 1. There exists a constant $K_\circ \in (0, \infty)$ such that

$$\|h\|^2 - \|g\|^2 \geqslant K_\circ\big(\|h\|_1 - \|g\|_1\big) \tag{9}$$

for all $h \geqslant g$ from $\mathcal{H}$, where $\|\cdot\|_1$ stands for the l_1-norm in $\mathbb{R}^n$, i. e.,

$$\|h\|_1 = \sum_{i=1}^n |h_i|.$$

Mention the following oracle inequality [16] for the exponential weighting of ordered smoothers in the case of white Gaussian noise with variance σ^2 that is diagonal Σ with $\sigma_{\min} = \sigma_{\max} = \sigma$.

Theorem 1. *Assume that $\mathcal{H}$ is a set of ordered multipliers, $\beta \geqslant 4$, and Condition 1 holds. Then, uniformly in $\theta \in R^n$,*

$$\mathsf{E}_\theta\|\bar\theta - \theta\|^2 \leqslant r^{\mathcal{H}}(\theta) + 2\beta\sigma^2 \log\left[C\left(1 + \frac{r^{\mathcal{H}}(\theta)}{\sigma^2}\right)\right].$$

This oracle inequality outperforms (in the form of the remainder term) Kniep's oracle inequality [17].

Theorem 2. *Uniformly in $\theta \in R^n$,*

$$\mathsf{E}_\theta\|\hat h \cdot Y - \theta\|^2 \leqslant r^{\mathcal{H}}(\theta) + K\sigma^2\sqrt{1 + \frac{r^{\mathcal{H}}(\theta)}{\sigma^2}}$$

where a minimizer of the unbiased risk estimate $\hat h = \operatorname{argmin}_{h\in\mathcal{H}} \bar r(Y, \hat\theta^h)$ corresponds to the case $\beta \to 0$ in exponential weighting and K is the generic constant.

The main result of this paper is the following new oracle inequality with remainder term of the same form as in [16] for the exponential weighting in the case of colored noise problem.

Theorem 3. *Assume that $\mathcal{H}$ is a set of ordered multipliers, $\beta \geqslant 4$, and Condition 1 holds. Then, uniformly in $\theta \in R^n$,*

$$\mathsf{E}_\theta\|\bar\theta - \theta\|^2 \leqslant r^{\mathcal{H}}(\theta) + 2\beta\sigma_{\max}^2 \log\left[C\left(1 + \frac{r^{\mathcal{H}}(\theta)}{\sigma_{\min}^2}\right)\right].$$

Here and in what follows, $C = C(C_\circ, K_\circ, \beta, \varkappa)$ denotes strictly positive and bounded constant depending on $C_\circ, K_\circ, \beta$, and $\varkappa$, where $\varkappa = \sigma_{\max}/\sigma_{\min}$.

For the case of stationary noise ξ with variance σ^2, one has $\sigma^2_{\min} = \sigma^2_{\max} = \sigma^2$ and the following

Corollary 1. *Assume that $\mathcal{H}$ is a set of ordered multipliers, $\beta \geqslant 4$, and Condition 1 holds. Then, uniformly in $\theta \in R^n$,*

$$\mathsf{E}_\theta \|\bar{\theta} - \theta\|^2 \leqslant r^{\mathcal{H}}(\theta) + 2\beta\sigma^2 \log\left[C\left(1 + \frac{r^{\mathcal{H}}(\theta)}{\sigma^2}\right)\right].$$

Here and in what follows, $C = C(C_\circ, K_\circ, \beta)$ denotes strictly positive and bounded constants depending on $C_\circ, K_\circ$, and β.

5 Simulations

To find out what value of β is good from a practical viewpoint and to compare the cases of white and coloured Gaussian noise, a numerical experiment has been carried out. The present author compares the exponential weighting methods applied to the set of cubic smoothing splines (as ordered smoothers) for $\beta = \{0, 1, 2, 4\}$ and for the equidistant design:

$$\mathcal{H} = \left\{h : h_k = \frac{1}{1 + [\alpha(k-1)]^4},\ \alpha > 0\right\}$$

where an asymptotic formula for the eigenvalues of Demmler–Reinsch basis was used in the case of equidistant design: $\lambda_k \asymp (\pi k)^4,\ k \to \infty$.

The scheme of the experiment is the following. For a given $A \in [0, 300]$, 100 000 replications of the observations

$$Y_k = \theta_k(A) + \xi_k,\ k = 1, \ldots, 400,$$

are generated. Here, $\theta(A) \in R^{400}$ is the Gaussian vector with independent components and

$$\mathsf{E}\theta_k(A) = 0, \quad \mathsf{E}\theta_k^2(A) = A \exp\left(-\frac{k^2}{2\Omega^2}\right)$$

where $\Omega = 50$.

Two types of the noise ξ were considered:

1) standard Gaussian white noise ($\sigma_i = 1$); and
2) Gaussian vector with covariance matrix Σ with eigenvalues $\sigma_i = i/400$, $i = 1, \ldots, 400$.

Next, the mean oracle risk

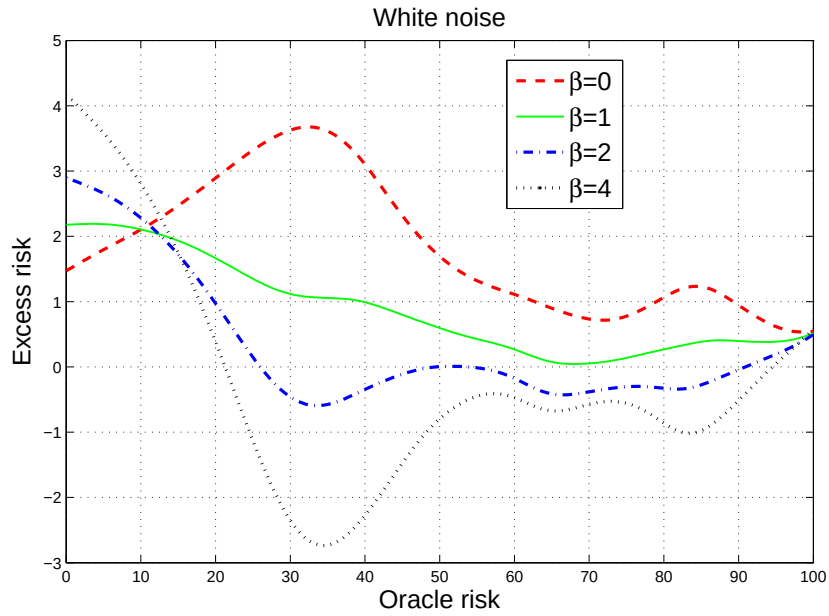
$$\bar{r}^{\mathcal{H}}(A) = \mathsf{E} \min_{h \in \mathcal{H}} \left\{ \|(1 - h) \cdot \theta(A)\|^2 + \|\sigma \cdot h\|^2 \right\}$$

and the mean excess risk

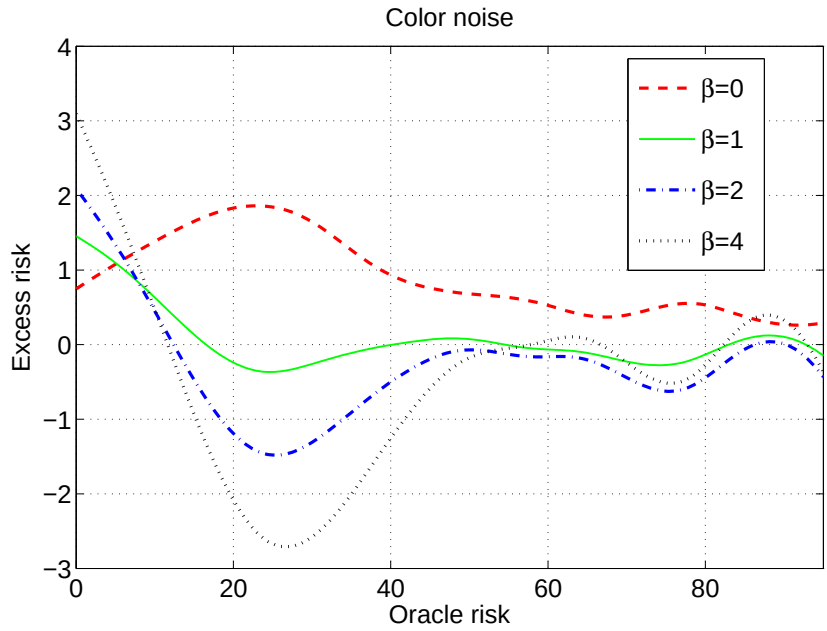
$$\bar{\Delta}_\beta(A) = \mathsf{E}\|\theta(A) - \bar{\theta}(Y)\|^2 - \bar{r}^{\mathcal{H}}(A)$$

were computed with the help of the Monte-Carlo method. Finally, the data $\{\bar{r}^{\mathcal{H}}(A),\ \bar{\Delta}_\beta(A),\ A \in [0, 300]\}$ are plotted in Fig. 1 to illustrate graphically the remainder term $\Delta_\beta(r^{\mathcal{H}}) = \mathsf{E}_\theta\|\bar{\theta} - \theta\|^2 - r^{\mathcal{H}}(\theta)$.

Looking at Fig. 1, one sees that there is no universal β minimizing the excess risk uniformly in θ. However, intuitively, it seems that a reasonable choice is $\beta \approx 1$ [15] but unfortunately, good oracle inequalities are not available for this case. Almost all methods demonstrate similar



(a)



(b)

Figure 1 Exponential weighting for the white (a) and colored (b) noise cases. The data $\{\bar{r}^{\mathcal{H}}(A), \bar{\Delta}_{\beta}(A),\ A \in [0, 300]\}$ that is the dependancy of excess risk on oracle risk is in the pictures

statistical performance (in Fig. 1, for the values of oracle risk bigger than 50). However, when $r^{\mathcal{H}}(\theta)/\sigma^2$ is not large, the exponential weighting works usually better (in Fig. 1, for the values of oracle risk from approximately 10 to 50).

6 Proofs

The main steps of the proof are based on a combination of methods for deriving oracle inequalities proposed in [16, 20]. Here, the main steps in the proof are sketched, all details are given below.

With the help of Stein's formula for the unbiased risk estimate, it can be shown that for $\beta \geqslant 4$,

$$\mathsf{E}_\theta \|\bar\theta - \theta\|^2 \leqslant \mathsf{E}_\theta \sum_{h\in\mathcal{H}} w^h(Y)\bar r(Y,\hat\theta^h) \leqslant r^{\mathcal{H}}(\theta) + 2\beta\sigma^2_{\max}\mathsf{E}_\theta \sum_{h\in\mathcal{H}} w^h(Y)\log\frac{\pi^h}{w^h(Y)} - 2\beta\sigma^2_{\max}\mathsf{E}_\theta \log\left\{\sum_{h\in\mathcal{H}}\pi^h \exp\left[-\frac{\bar r(Y,\hat\theta^h)-\bar r(Y,\hat\theta^{\hat h})}{2\beta\sigma^2_{\max}}\right]\right\} \quad (10)$$

where $\hat h$ is the minimizer of the unbiased risk estimate $\hat h = \arg\min_{h\in\mathcal{H}} \bar r(Y,\hat\theta^h)$.

To control the right-hand side at this equation, make use of the ordering property of estimates $\hat\theta^h,\ h\in\mathcal{H}$. First, check that if π^h is defined by (8), then

$$\sum_{h\in\mathcal{H}}\pi^h \exp\left[-\frac{\bar r(Y,\hat\theta^h)-\bar r(Y,\hat\theta^{\hat h})}{2\beta\sigma^2_{\max}}\right] \geqslant \sum_{h\geqslant\hat h}\pi^h \exp\left[-\frac{\bar r(Y,\hat\theta^h)-\bar r(Y,\hat\theta^{\hat h})}{2\beta\sigma^2_{\max}}\right] \geqslant 1$$

and so, the last term in Eq. (10) is always negative.

The most difficult and delicate part of the proof is related to the average Kullback–Leibler divergence $\mathsf{E}_\theta\sum_{h\in\mathcal{H}} w^h(Y)\log(w^h(Y)/\pi^h)$. To compute a good lower bound for this value, follow the approach proposed in [20]. The main idea here is to make use of the following property of the unbiased risk estimate: for any sufficiently small $\varepsilon<1$, there exists $\hat h^\varepsilon$ depending on Y such that with probability 1, for all $h\geqslant\hat h^\varepsilon$,

$$\bar r(Y,\hat\theta^h) - \bar r(Y,\hat\theta^{\hat h}) \geqslant 2\beta\varepsilon\big[\|\sigma\cdot h\|^2 - \|\sigma\cdot\hat h\|^2\big] + 2\beta\sigma^2_{\min}\,.$$

This equation means that $w^h(Y)$ are exponentially decreasing for large h. With this property, one obtains the following entropy bound:

$$\sum_{h\in\mathcal{H}} w^h(Y)\log\frac{\pi^h}{w^h(Y)} \leqslant \log\left[\sum_{h\leqslant\hat h^\varepsilon}\pi^h + \frac{C}{\varepsilon}\exp\left(\frac{C}{\varepsilon}\right)\right].$$

The rest of the proof consists in deriving the following bound from (9) and (8):

$$\sum_{h\leqslant\hat h^\varepsilon}\pi^h \leqslant 1 + \frac{\|\sigma\cdot\hat h^\varepsilon\|^2}{K_\circ\beta\sigma^2_{\max}}$$

and

$$\sqrt{\mathsf{E}_\theta\|\sigma\cdot\hat h^\varepsilon\|^2} \leqslant \sqrt{\frac{r^{\mathcal{H}}(\theta)}{1-2\beta\varepsilon}} + \frac{\sqrt{1+2\beta}}{1-2\beta\varepsilon}\frac{\sqrt{KC_\circ}}{\sigma^2_{\min}}.$$

Finally, combining the above equations, one arrives at (1).

7 Concluding Remarks

Based on the probabilistic properties of the unbiased risk estimate, the oracle inequality was proved for the method of aggregation of smoothing splines for the regression estimation problem in the case of colored noise. However, it seems that no good oracle inequalities are available for the reasonable choice of β parameter in the definition of aggregating weights. Numerical results demonstrate similar statistical performance for different choice of β parameter.

References

[1] Engl, H. W., M. Hanke, and A. Neubauer. 1996. *Regularization of inverse problems. Mathematics and its applications*. Dordrecht: Kluwer Academic Publishers Group. 375 p.

[2] Speckman, P. 1985. Spline smoothing and optimal rates of convergence in nonparametric regression. *Ann. Statist.* 13:970–983.

[3] Green, P. J., and B. W. Silverman. 1994. *Nonparametric regression and generalized linear models. A roughness penalty approach*. Chapman and Hall. 184 p.

[4] Tikhonov, A. N., and V. A. Arsenin. 1977. *Solution of ill-posed problems*. Scripta ser. in mathematics. Washington, DC – New York, NY: V. H. Winston & Sons – John Wiley & Sons. 258 p.

[5] Stein, C. 1981. Estimation of the mean of a multivariate normal distribution. *Ann. Stat.* 9:1135–1151.

[6] Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. *2nd Symposium (International) on Information Theory Proceedings*. 267–281.

[7] Mallows, C. L. 1973. Some comments on C_p. *Technometrics* 15:661–675.

[8] Nemirovski, A. 2000. *Topics in non-parametric statistics*. Lectures notes in mathematics ser. Berlin: Springer-Verlag. 197 p.

[9] Catoni, O. 2004. *Statistical learning theory and stochastic optimization*. Lectures notes in mathematics ser. Berlin: Springer-Verlag. 279 p.

[10] Yang, Y. 2004. Aggregating regression procedures to improve performance. *Bernoulli* 10:25–47.

[11] Lecué, G. 2007. Simultaneous adaptation to the margin and to complexity in classification. *Ann. Stat.* 35:1698–1721.

[12] Rigollet, P., and A. B. Tsybakov. 2007. Linear and convex aggregation of density estimators. *Math. Methods Statist.* 16:260–280.

[13] Rigollet, Ph., and A. Tsybakov. 2011. Sparse estimation by exponential weighting. arXiv:1108.5116v1 [math.ST].

[14] Leung, G., and A. Barron. 2006. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* 52(8):3396–3410.

[15] Dalayan, A., and J. Salmon. 2011. Sharp oracle inequalities for aggregation of affine estimators. arXiv:1104.3969v2 [math.ST].

[16] Chernousova, E., Yu. Golubev, and E. Krymova. 2013. Ordered smoothers with exponential weighting. *Electron. J. Statist.* 7.

[17] Kneip, A. 1994. Ordered linear smoothers. *Ann. Stat.* 22:835–866.

[18] Wahba, G. 1990. *Spline models for observational data*. Philadelphia, PA: SIAM. 161 p.

[19] Demmler, A., and C. Reinsch. 1975. Oscillation matrices with spline smoothing. *Numerische Mathematik* 24:375–382.

[20] Golubev, Yu. 2012. Exponential weighting and oracle inequalities for projection methods. *Problems Inform. Transmission* 3. arXiv:1206.4285.

Received June 15, 2015

Литература

[1] *Engl H. W., Hanke M., Neubauer A.* Regularization of inverse problems. Mathematics and its applications. — Dordrecht: Kluwer Academic Publishers Group, 1996. 375 p.

[2] *Speckman P.* Spline smoothing and optimal rates of convergence in nonparametric regression // Ann. Stat., 1985. No. 13. P. 970–983.

[3] *Green P. J., Silverman B. W.* Nonparametric regression and generalized linear models. A roughness penalty approach. — Chapman and Hall, 1994. 184 p.

[4] *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М.: Наука, 1979. 285 с.

[5] *Stein C.* Estimation of the mean of a multivariate normal distribution // Ann. Stat., 1981. No. 9. P. 1135–1151.

[6] *Akaike H.* Information theory and an extension of the maximum likelihood principle // 2nd Symposium (Intenational) on Information Theory Proceedings, 1973. P. 267–281.

[7] *Mallows C. L.* Some comments on C_p // Technometrics, 1973. No. 15. P. 661–675.

[8] *Nemirovski A.* Topics in non-parametric statistics. — Lectures notes in matematics ser. — Berlin: Springer-Verlag, 2000. 197 p.

[9] *Catoni O.* Statistical learning theory and stochastic optimization. — Lectures notes in mathematics ser. — Berlin: Springer-Verlag, 2004. 279 p.

[10] *Yang Y.* Aggregating regression procedures to improve performance // Bernoulli, 2004. No. 10. P. 25–47.

[11] *Lecué G.* Simultaneous adaptation to the margin and to complexity in classification // Ann. Stat., 2007. No. 35. P. 1698–1721.

[12] *Rigollet P., Tsybakov A. B.* Linear and convex aggregation of density estimators // Math. Methods Statist., 2007. No. 16. P. 260–280.

[13] *Rigollet Ph., Tsybakov A.* Sparse estimation by exponential weighting. 2011. arXiv:1108.5116v1 [math.ST].

[14] *Leung G., Barron A.* Information theory and mixing least-squares regressions // IEEE Trans. Inform. Theory, 2006. Vol. 52. No. 8. P. 3396–3410.

[15] *Dalayan A., Salmon J.* Sharp oracle inequalities for aggregation of affine estimators. 2011. arXiv:1104.3969v2 [math.ST].

[16] *Chernousova E., Golubev Yu., Krymova E.* Ordered smoothers with exponential weighting // Electron. J. Statist., 2013. No. 7.

[17] *Kneip A.* Ordered linear smoothers // Ann. Stat., 1994. No. 22. P. 835–866.

[18] *Wahba G.* Spline models for observational data. — Philadelphia, PA, USA: SIAM, 1990. 161 p.

[19] *Demmler A., Reinsch C.* Oscillation matrices with spline smoothing // Numerische Mathematik, 1975. No. 24. P. 375–382.

[20] *Голубев Г. К.* Экспоненциальное взвешивание и оракульные неравенства для проекционных оценок // Пробл. передачи информ., 2012. Т. 48. № 3. С. 83–95.

Поступила в редакцию 15.06.2015

Нелинейная многомерная фильтрация многокомпонентных изображений*

Е. В. Медведева, И. С. Трубин, Е. А. Устюжанина, А. В. Лалетин
emedv@mail.ru
Вятский государственный университет, Киров, Россия

Предложен метод нелинейной многомерной фильтрации многокомпонентных изображений, искаженных аддитивным белым гауссовским шумом. Повышение качества зашумленных изображений обеспечивается за счет эффективного использования статистической избыточности многокомпонентных изображений. Рассмотрен частный случай многокомпонентных изображений — цветные RGB изображения, каждая из цветовых компонент которого представляет собой g-разрядное цифровое полутоновое изображение (ЦПИ). Метод основан на представлении многокомпонентных g-разрядных ЦПИ набором разрядных двоичных изображений (РДИ), аппроксимации их трехмерной цепью Маркова и применении теории фильтрации условных марковских процессов. Предложено улучшить качество восстановленных изображений за счет повышения точности вычисления статистических характеристик для каждой локальной области внутри изображений и между цветовыми компонентами. Для оценки статистических характеристик использовано скользящее окно. Приведены результаты моделирования, подтверждающие эффективность разработанного метода. **Ключевые слова**: *многокомпонентные изображения; многомерная нелинейная фильтрация; многомерные цепи Маркова; разрядные двоичные изображения; статистическая избыточность изображений*

DOI: 10.21469/22233792.1.13.02

Multidimensional nonlinear filtration of multicomponent images*

E. V. Medvedeva, I. S. Trubin, E. A. Ustyuzhanina, and A. V. Laletin
Vyatka State University, 36 Moskovskaya st., Kirov, Russia

The goal of this paper is to develop a method of nonlinear multidimensional multicomponent images filtering based on mathematical apparatus of Markov chains. The method allows efficient use of the statistical redundancy of the image to improve the quality of image distorted by white Gaussian noise. Multidimensional signals of multicomponent images have a much greater statistical redundancy than single image. This redundancy would be appropriate for use to improve the quality of the restoration of noisy images. Special cases of multicomponent images are RGB image, each color component of which is a g-bit half-tone digital image (HTDI). The nature of the statistical relationship between elements within the HTDI and among the elements of color components (RG, GB, BR) allows one to use this method as an approximation for the three-dimensional (3D) color images of a Markov chain with several states and for bit binary image (bit planes) of two color components of the 3D Markov chain with two states. This approximation makes it possible to apply the theory of filtration of conditional Markov

*Работа выполнена в рамках базовой части государственного задания в сфере научной деятельности по заданию № 2014/61.

processes for the development of filtering method of multicomponent images. Realistic images contain the regions with varying degrees of detail and different statistical characteristics. The authors propose improving the accuracy of calculation of the statistical characteristics of each local region within the image and between the color components to improve the quality of the reconstructed image. A sliding window has been used to estimate local statistical characteristics of the image. The authors present the results of modeling of the suggested algorithm of a 3D nonlinear filtration with use of the sliding window and earlier developed algorithm of a two-dimensional filtration of color (RGB) images. The developed 3D filter taking into account the sliding window provided to reducing quantity of the artifacts similar to influence of pulse hindrances to provide allocation of borders and small-sized objects more exact. The gain in the mean square error is from 30% to 70%, respectively, in the range of the signal/noise relations $\rho^2_{\text{in}} = -9 \ldots -3$ dB.

Keywords: *multicomponent images; nonlinear multidimensional filtering; multidimensional Markov chains; binary g-bit digital images; statistical redundancy of the image*

DOI: 10.21469/22233792.1.13.02

1 Введение

Для ряда современных видеосистем характерно использование многокомпонентных изображений. Примером являются мульти- и гиперспектральные системы дистанционного зондирования. Изображения, полученные такими системами, содержат десятки и даже сотни спектральных каналов, в которых присутствуют помехи различной интенсивности. Для восстановления изображений на фоне помех с целью решения дальнейших задач обработки изображений: выделения объектов и оценки их параметров, классификации, распознавания и т. д., — используют фильтрацию. К настоящему времени разработано много разнообразных алгоритмов фильтрации [1–5], синтезируемых для конкретной модели помех. Так, например, известные линейные алгоритмы фильтрации, основанные на применении локальных операторов [2, 3], эффективны при больших отношения сигнал/шум, но с увеличением мощности шума приводят к сглаживанию мелких деталей и размытию границ объектов. Из нелинейных методов фильтрации, в силу малых вычислительных затрат, наибольшее распространение получили методы, основанные на различных модификациях медианной фильтрации, фильтры окрестных элементов, включая билатеральные фильтры и фильтры нелокальных значений [2–4]. Недостатком перечисленных фильтров, в малой степени искажающих резкие границы изображений и хорошо подавляющих импульсные помехи, является низкая эффективность при наличии белого гауссовского шума (БГШ). На настоящий момент наиболее эффективными фильтрами при наличии БГШ являются фильтры BM3D и BM4D (block-matching and 3D/4D filtering) [1, 5], объединяющие несколько дополняющих друг друга механизмов, один из которых связан с фильтрацией на основе дискретного косинусного преобразования в группах найденных подобных блоков. Основным недостатком работы фильтров BM3D и BM4D являются низкое быстродействие и размытие границ при малых отношениях сигнал/шум.

Следует также отметить, что большинство известных алгоритмов фильтрации являются двумерными, применяются к каждой отдельной компоненте изображения и, как следствие, не всегда обеспечивают надлежащее качество изображения, особенно в условиях действия шумов большой интенсивности.

В свою очередь многокомпонентные изображения представляют собой многомерные сигналы и обладают значительно большей статистической избыточностью, чем однокомпонентные изображения, которую целесообразно использовать для повышения качества

восстановления зашумленных изображений. Поэтому разработка алгоритмов фильтрации многокомпонентных изображений, эффективно использующих статистическую избыточность изображений и тем самым позволяющих повысить качество их восстановления, является актуальной задачей.

Частным случаем многокомпонентных изображений можно считать цветные RGB изображения, каждая из цветовых компонент которого представляет собой g-разрядное ЦПИ. Известно, что между отдельными областями ЦПИ, принадлежащих разным цветовым компонентам, существует большая статистическая зависимость между элементами изображения. Например, области желтого цвета одинаково хорошо выделены на красной и зеленой компонентах, а области белого цвета — на всех трех компонентах. Таким образом, учитывая характер статистической связи между элементами внутри ЦПИ и между элементами цветовых компонент (RG, GB, BR), можно предположить, что цветные RGB изображения допускают аппроксимацию трехмерной цепью Маркова с несколькими состояниями.

Цель предлагаемой работы — разработка метода нелинейной многомерной фильтрации многокомпонентных изображений на основе математического аппарата цепей Маркова и эффективного использования статистической избыточности, позволяющего повысить качество изображений, искаженных БГШ.

2 Математическая модель RGB изображения

При обработке ЦПИ с числом уровней яркости 2^g возникает проблема хранения в памяти и оперирования с матрицами вероятностей переходов размерностью $2^g \times 2^g$. Такая обработка ЦПИ требует больших вычислительных ресурсов. В работах [6–10] предложено ЦПИ, представленные g-разрядными двоичными числами, разбивать на g РДИ или битовых плоскостей, что позволило снизить вычислительные ресурсы за счет оперирования с матрицами вероятностей переходов размером 2×2.

На рис. 1 представлен график усредненной зависимости вероятностей переходов между элементами трех цветовых компонент (RG, GB, BR) от номера разряда ЦПИ, подтверждающий, что между элементами RGB изображения, особенно предлежащими старшим разрядам ЦПИ, существует большая статистическая зависимость.

Если l-е РДИ, $l = 1, \ldots, g$, представляет собой марковское случайное поле с разделимой автокорреляционной функцией, то в этом случае l-е РДИ цветовых компонент можно

Рис. 1 Усредненная зависимость вероятностей переходов между элементами трех цветовых компонент от номера разряда ЦПИ

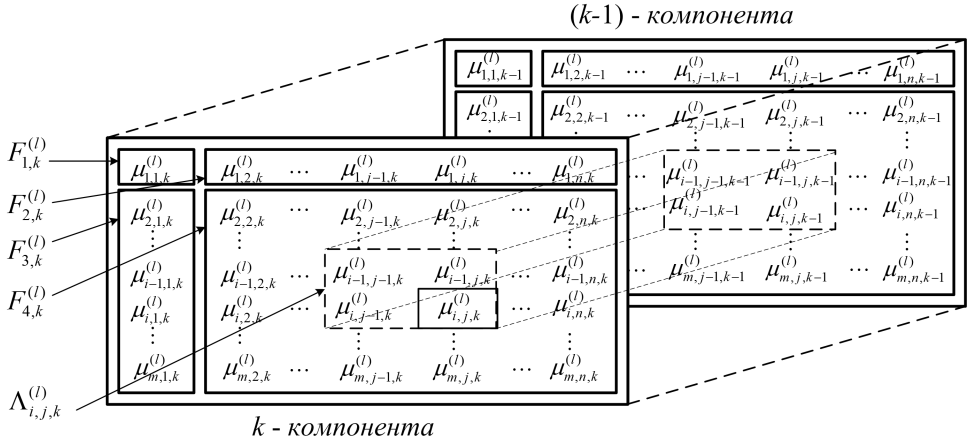


Рис. 2 Разрядные двоичные изображения двух цветовых компонент l-го разряда ЦПИ

представить суперпозицией трех одномерных цепей Маркова по горизонтали, вертикали и между компонентами с двумя равновероятными состояниями $M_1^{(l)}, M_2^{(l)}$ и матрицами вероятностей переходов по горизонтали ${}^1\Pi = \|{}^1\pi_{ij}^{(l)}\|_{2\times 2}$, вертикали ${}^2\Pi = \|{}^2\pi_{ij}^{(l)}\|_{2\times 2}$ и ${}^4\Pi = \|{}^4\pi_{ij}^{(l)}\|_{2\times 2}$ между цветовыми компонентами (RG, GB, BR).

На рис. 2 показаны РДИ двух цветовых компонент l-го разряда ЦПИ, разделенных на области $F_i^{(l)}$ $(i = 1, \ldots, 4)$, элементы которых являются цепью Маркова различной размерности. Состояние элемента $\nu_4^{(l)} = \mu_{i,j,k}^{(l)}$ области $F_4^{(l)}$ зависит от состояния семи соседних элементов, входящих в его окрестность, где $\nu_1^{(l)} = \mu_{i,j-1,k}^{(l)}, \nu_2^{(l)} = \mu_{i-1,j,k}^{(l)}, \nu_3^{(l)} = \mu_{i-1,j-1,k}^{(l)}, \nu_1'^{(l)} = \mu_{i,j-1,k-1}^{(l)}, \nu_2'^{(l)} = \mu_{i-1,j,k-1}^{(l)}, \nu_3'^{(l)} = \mu_{i-1,j-1,k-1}^{(l)}, \nu_4'^{(l)} = \mu_{i,j,k-1}^{(l)}$ (i, j — пространственные координаты; $k = 1, 2, 3$ — номер цветовой компоненты для RGB изображения). Для построения математической модели l-го РДИ цветовых компонент потребуется семь матриц вероятностей переходов: три — основные ${}^1\Pi, {}^2\Pi, {}^4\Pi$ и четыре — дополнительные, полученные на основе трех априорно заданных: ${}^3\Pi^{(l)} = {}^1\Pi^{(l)} \times {}^2\Pi^{(l)}$; ${}^5\Pi^{(l)} = {}^1\Pi^{(l)} \times {}^4\Pi^{(l)}$; ${}^6\Pi^{(l)} = {}^2\Pi^{(l)} \times {}^4\Pi^{(l)}$; ${}^7\Pi^{(l)} = {}^3\Pi^{(l)} \times {}^4\Pi^{(l)}$, определяющих статистическую связь элементов $\nu_3^{(l)}, \nu_1'^{(l)}, \nu_2'^{(l)}, \nu_3'^{(l)}$, с элементом $\nu_4^{(l)}$ соответственно [6, 8].

3 Метод нелинейной многомерной фильтрации изображений

Предполагалось, что двоичные символы разрядов ЦПИ цветовых компонент передаются бинарными импульсными сигналами по радиоканалу независимо друг от друга в присутствии аддитивного БГШ $n(t)$ с нулевым средним и дисперсией σ_n^2.

На основе трехмерной математической модели и результатов, полученных в работах [6, 8], синтезировано уравнение для апостериорной вероятности состояний элемента $\nu_4^{(l)} = \mu_{i,j,k}^{(l)}$ (см. рис. 2), выраженное через одномерные апостериорные вероятности $p(\nu_i^{(l)})$ и вероятности перехода состояний элементов окрестности $\Lambda_{i,j,k} = \left\{\nu_1^{(l)}, \nu_2^{(l)}, \nu_3^{(l)}, \nu_1'^{(l)}, \nu_2'^{(l)}, \nu_3'^{(l)}, \nu_4'^{(l)}\right\}$ к состоянию элемента $\nu_4^{(l)}$:

$$p_j(\nu_4^{(l)}) = c \exp\{f\left(M_j(\nu_4^{(l)})\right)\} \frac{p(\nu_1^{(l)})^1\pi_{ij}^{(l)} p(\nu_2^{(l)})^2\pi_{ij}^{(l)} p(\nu_4'^{(l)})^4\pi_{ij}^{(l)} p(\nu_3'^{(l)})^7\pi_{ij}^{(l)}}{p(\nu_3^{(l)})^3\pi_{ij}^{(l)} p(\nu_1'^{(l)})^5\pi_{ij}^{(l)} p(\nu_2'^{(l)})^6\pi_{ij}^{(l)}}, \qquad (1)$$

где c — коэффициент нормировки; $f\left(M_j(\nu_4^{(l)})\right)$ — логарифм функции правдоподобия элемента $\nu_4^{(l)}$ l-го РДИ в k-й компоненте.

Разделив уравнение (1) при $j = 1$ на уравнение при $j = 2$ и прологарифмировав слева и справа, получим рекуррентное уравнение трехмерной нелинейной фильтрации элементов l-го РДИ вида [6, 8]:

$$u\left(\nu_4^{(l)}\right) = \left[f\left(M_1\left(\nu_4^{(l)}\right)\right) - f\left(M_2\left(\nu_4^{(l)}\right)\right)\right] + u\left(\nu_1^{(l)}\right) + z_1\left[u\left(\nu_1^{(l)}\right), {}^1\pi_{ij}^{(l)}\right] + \\ + u\left(\nu_2^{(l)}\right) + z_2\left[u\left(\nu_2^{(l)}\right), {}^2\pi_{ij}^{(l)}\right] + u\left(\nu_4'^{(l)}\right) + z_4\left[u\left(\nu_4'^{(l)}\right), {}^4\pi_{ij}^{(l)}\right] + \\ + u\left(\nu_3'^{(l)}\right) + z_7\left[u\left(\nu_3'^{(l)}\right), {}^7\pi_{ij}^{(l)}\right] - u\left(\nu_3^{(l)}\right) - z_3\left[u\left(\nu_3^{(l)}\right), {}^3\pi_{ij}^{(l)}\right] - \\ - u\left(\nu_1'^{(l)}\right) - z_5\left[u\left(\nu_1'^{(l)}\right), {}^5\pi_{ij}^{(l)}\right] - u\left(\nu_2'^{(l)}\right) - z_6\left[u\left(\nu_2'^{(l)}\right), {}^6\pi_{ij}^{(l)}\right] \geqslant H, \quad (2)$$

где $u(\nu_4^{(l)}) = \ln\left(p_1\left(\nu_4^{(l)}\right)/p_2\left(\nu_4^{(l)}\right)\right)$ — логарифм отношения апостериорных вероятностей состояния фильтруемого элемента $\nu_4^{(l)}$ l-го РДИ; $\left[f\left(M_1(\nu_4^{(l)})\right) - f\left(M_2(\nu_4^{(l)})\right)\right] = 4\rho_{\text{in}}^2\left[\pm 1 + \xi/(\sqrt{2}\rho_{\text{in}})\right]$ — разность логарифмов функций правдоподобия на выходе фазового дискриминатора при $M_1 = 1$, $M_2 = -1$; $\rho_{\text{in}}^2 = A^2T/N_0$ — отношение сигнал/шум по мощности (A — амплитуда; T — длительность импульса; N_0 — спектральная плотность мощности шума); ξ — шум в k-м такте; H — порог, выбранный в соответствии с критерием идеального наблюдателя (для данного алгоритма $H = 0$);

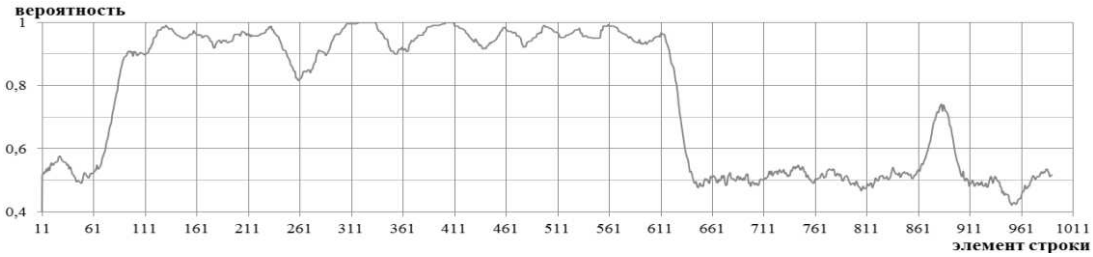
$$z_r(\cdot) = \ln\frac{{}^r\pi_{ii}^{(l)} + {}^r\pi_{ji}^{(l)}\exp\left(-u\left(\nu_r^{(l)}\right)\right)}{{}^r\pi_{jj}^{(l)} + {}^r\pi_{ji}^{(l)}\exp\left(u\left(\nu_r^{(l)}\right)\right)}, \quad r = 1, \ldots, 7\,.$$

В нелинейной функции $z_r(\cdot)$ содержится вся априорная информация о степени корреляции между элементами изображения. Соответственно эффективность фильтрации непосредственно будет зависеть от точности вычисленных оценок элементов матриц вероятностей переходов.

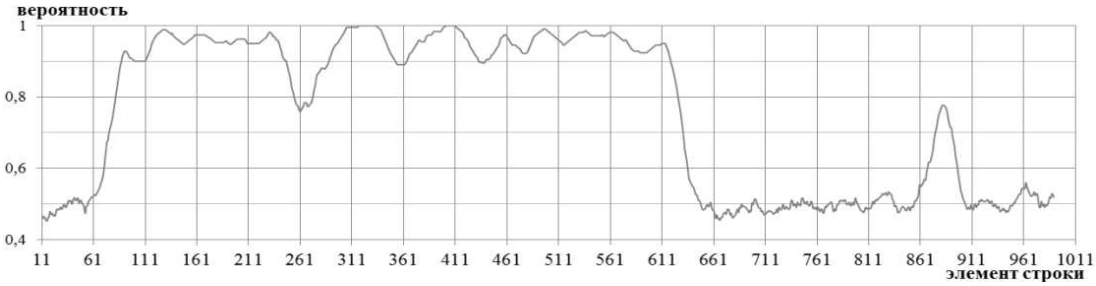
В работах [6, 8] алгоритм нелинейной фильтрации (2) использовался для восстановления зашумленных видеопоследовательностей, представляющих собой трехмерные сигналы. Проведенные исследования показали, что в целом алгоритм нелинейной фильтрации видеоизображений (2), разрушенных БГШ, повышает качество их восстановления за счет эффективного использования статистической избыточности. Однако в уравнение (2) подставлялись усредненные вероятности переходов, вычисленные по всему РДИ, и не учитывалось, что реальные изображения содержат области с разной степенью детальности и различными статистическими характеристиками. Также следует учесть, что некоторые области на l-х РДИ, принадлежащие разным цветовым компонентам, слабо коррелированы. Поэтому улучшить качество восстановленных изображений можно за счет вычисления оценок элементов матриц вероятностей переходов для каждой локальной области и последующей подстановки их в уравнение фильтрации (2). В данной работе, учитывая локальные изменения статистических характеристик на многокомпонентных изображениях, для их вычислений предлагается использовать метод «скользящего окна».

В пределах скользящего окна вычисляются оценки вероятностей переходов по горизонтали ${}^1\hat{\pi}_{ij}^{(l)}$, вертикали ${}^2\hat{\pi}_{ij}^{(l)}$ и между двумя цветовыми компонентами ${}^4\hat{\pi}_{ij}^{(l)}$ l-го РДИ. Вычисленные оценки вероятностей переходов подставляются в уравнение фильтрации (2), и восстанавливается элемент, соответствующий центральному элементу окна. В данной работе для оценки эффективности оптимальной фильтрации статистические характеристики вычислялись по исходному, незашумленному изображению.

($а$) Фрагмент искусственного РДИ



($б$) Оценка вероятности перехода по горизонтали ${}^1\hat{\pi}_{ij}^{(l)}$



($в$) Оценка вероятности перехода по вертикали ${}^2\hat{\pi}_{ij}^{(l)}$

Рис. 3 Пример вычисления оценок ${}^1\hat{\pi}_{ij}^{(l)}$ и ${}^2\hat{\pi}_{ij}^{(l)}$ искусственного РДИ в пределах скользящего окна

На рис. 3 приведен пример вычисленных оценок ${}^1\hat{\pi}_{ij}^{(l)}$ и ${}^2\hat{\pi}_{ij}^{(l)}$ для одной строки искусственного РДИ в пределах скользящего окна размером 21×21. Искусственное РДИ содержит области с разными статистическими характеристиками и получено по двумерной математической модели [7] с использованием матриц вероятностей переходов для каждой локальной области:

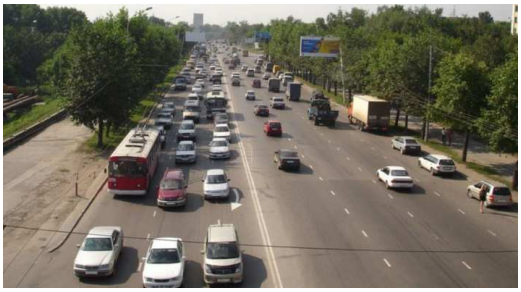
$$ {}^1\Pi = {}^2\Pi = \begin{Vmatrix} 0{,}9 & 0{,}1 \\ 0{,}1 & 0{,}9 \end{Vmatrix}; \quad {}^1\Pi = {}^2\Pi = \begin{Vmatrix} 0{,}5 & 0{,}5 \\ 0{,}5 & 0{,}5 \end{Vmatrix}. $$

Из анализа графиков (см. рис. 3) видно, что оценки вероятностей переходов в пределах окна совпадают с истинными вероятностями (2) для каждой текстурной области. Очевидно, чем больше размер локальных областей с однородными статистическими характеристиками, тем большее по размеру окно следует использовать. В то же время применение окон больших размеров для небольших локальных областей приведет к усреднению статистических характеристик в пределах окна и увеличению погрешности оценок.

4 Результаты моделирования

Моделирование выполнялось на реальных цветных (RGB) изображениях различной размерности при разных отношениях сигнал/шум в элементе l-го РДИ ρ_{in}^2 на входе устройства фильтрации. Отношение сигнал/шум по мощности ρ_{in}^2 априорно принято одинаковым для всех g РДИ.

На рис. 4 показано сравнение результатов обработки реального цветного изображения размером 970×534 алгоритмом двумерной фильтрации без учета скользящего окна [6, 8]



(*а*) Исходное изображение

(*б*) Фрагмент исходного изображения

(*в*) Зашумленное изображение ($\rho^2_{\text{in}} = -6$) дБ

(*г*) Восстановленное двумерным фильтром

(*д*) Восстановленное трехмерным фильтром с использованием сканирующего окна

Рис. 4 Нелинейная фильтрация цветного изображения разными методами

и разработанным алгоритмом трехмерной фильтрации с учетом скользящего окна. На рис. 4, *а* приведено исходное тестовое изображение. Далее показаны увеличенные фрагменты: (*б*) исходного изображения; (*в*) зашумленного БГШ изображения при $\rho^2_{\text{in}} = -6$ дБ; (*г*) восстановленного двумерным алгоритмом без учета скользящего окна (т. е. выполнена независимая фильтрация трех цветовых компонент); (*д*) трехмерным нелинейным фильтром с учетом скользящего окна.

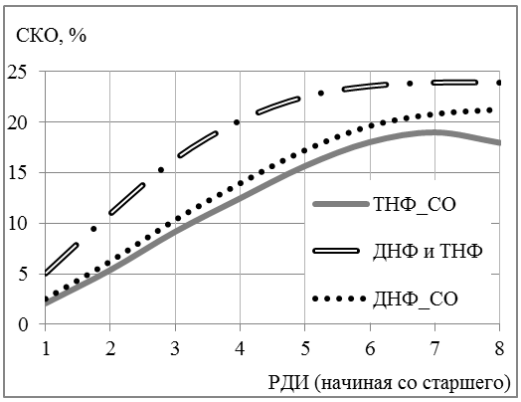


Рис. 5 Среднеквадратичная ошибка при двух- и трехмерной фильтрации ($\rho^2_{\text{in}} = -6$ дБ)

Из приведенных результатов видно, что трехмерный фильтр с учетом скользящего окна позволил уменьшить количество артефактов, подобных воздействию импульсных помех, обеспечить более точное выделение границ и малоразмерных объектов.

Для оценки качества изображения вычислялась побитовая среднеквадратичная ошибка (СКО):

$$\text{СКО} = \frac{1}{NMK}\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{M}\Big(x_{ik} - y_{jk}\Big)^2,$$

где x_{ik} и y_{jk} — исходное и восстановленное l-е РДИ k-й компоненты; M, N — размеры изображения; K — количество цветовых компонент.

На рис. 5 представлены зависимости СКО в тестовом изображении, восстановленном двумерными нелинейными фильтрами без учета (ДНФ) и с учетом скользящего окна (ДНФ_СО), а также трехмерными нелинейными фильтрами без учета (ТНФ) и с учетом скользящего окна (ТНФ_СО) от номера разряда ЦПИ при отношении сигнал/шум на входе приемного устройства $\rho^2_{\text{in}} = -6$ дБ. Двух- и трехмерные фильтры без учета скользящего окна дают близкие результаты в СКО, поэтому на графике они представлены одной линией. Трехмерный фильтр с учетом скользящего окна позволяет точнее вычислить статистические характеристики для каждой локальной области как внутри, так и между РДИ разных цветовых компонент. Следует отметить, что меньшее количество ошибок на-

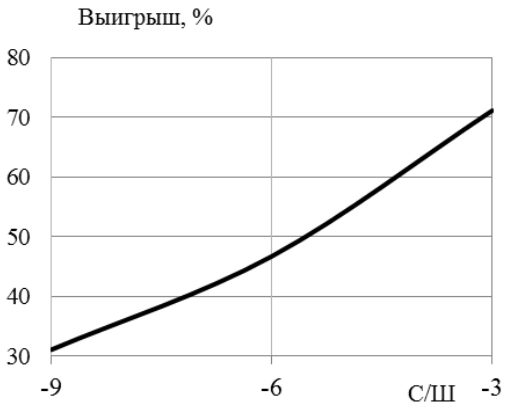


Рис. 6 Выигрыш в СКО для трехмерного фильтра с учетом окна

блюдается при восстановлении старших разрядов ЦПИ, что является весьма важным при повышении качества изображения.

На рис. 6 представлен выигрыш в СКО разработанным трехмерным алгоритмом фильтрации (ТНФ_СО) относительно двумерного алгоритма (ДНФ) при разных отношениях сигнал/шум. В диапазоне отношений сигнал/шум $\rho^2_{\text{in}} = -9, \ldots, -3$ дБ выигрыш в СКО составляет от 30% до 70% соответственно.

Для неизвестных статистических характеристик изображения необходимо применять адаптивные алгоритмы обработки, позволяющие непосредственно в процессе приема изображений вычислять оценки элементов матриц вероятностей переходов и выполнять адаптацию параметров алгоритма фильтрации многокомпонентных изображениях.

5 Заключение

Разработанный алгоритм многомерной нелинейной фильтрации с использованием скользящего окна, за счет повышения точности вычисления статистических характеристик для каждой локальной области и учета межкомпонентной избыточности, позволил точнее выделить объекты малоразмерной формы и контуры объектов и тем самым повысить качество многокомпонентных изображений, искаженных БГШ. Алгоритм эффективен при малых отношениях сигнал/шум. Дальнейшие исследования будут направлены на разработку адаптивного алгоритма фильтрации, в котором статистические характеристики будут вычисляться в пределах скользящего окна по зашумленному изображению.

Литература

[1] *Dabov K., Foi A., Katkovnik V., Egiazarian K.* Image denoising by sparse 3-D transform-domain collaborative filtering // IEEE Trans. Image Processing, 2007. Vol. 16. No. 8. P. 2080–2095.

[2] *Шовенгердт Р. А.* Дистанционное зондирование. Модели и методы обработки изображений. — М.: Техносфера, 2010. 594 с.

[3] *Гонсалес Р., Вудс Р.* Цифровая обработка изображений. — М.: Техносфера, 2012. 1104 с.

[4] *Самойлин Е. А.* Метод различия случайных сигналов многокомпонентных изображений и импульсных помех на основе свойства межканальной избыточности // Цифровая обработка сигналов, 2014. № 3. С. 2–8.

[5] *Возель Б., Кожемякин Р. А., Лукин В. В., Рубель А. С., Чобану М. К.* Предсказание эффективности фильтрации при обработке многоканальных изображений // Сб. научн. тр. 17-й Междунар. конф. «Цифровая обработка сигналов и ее применение». — М., 2015. С. 707–711.

[6] *Петров Е. П., Медведева Е. В., Метелев А. П.* Метод комбинированной нелинейной фильтрации коррелированных видеоизображений // Нелинейный мир, 2010. № 11. С. 677–684.

[7] *Петров Е. П., Медведева Е. В., Метелев А. П.* Метод синтеза математических моделей видеоизображений на основе многомерных цепей Маркова // Нелинейный мир, 2011. № 4. С. 213–231.

[8] *Petrov E. P., Trubin I. S., Medvedeva E. V., Smolskiy S. M.,* Development of nonlinear filtering algorithms of digital half-tone images // Integrated models for information communication systems and net-works: Design and development. — IGI Global, 2013. P. 278–304.

[9] *Лалетин А. В., Медведева Е. В., Устюжанина Е. А.* Метод двумерной нелинейной фильтрации изображений с использованием скользящего окна // Сб. докл. Всеросс. конф. с международным участием «Радиоэлектронные средства получения, обработки и визуализации информации». — Н. Новгород, 2014. С. 217–221.

[10] *Лалетин А. В., Медведева Е. В., Устюжанина Е. А.* Метод повышения качества видеоизображений, искаженных шумом // Сб. научн. тр. 17-й Междунар. конф. «Цифровая обработка сигналов и ее применение». — М., 2015. С. 715–719.

Поступила в редакцию 09.07.2015

References

[1] Dabov, K., A. Foi, V. Katkovnik, and K. Egiazarian. 2007. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. Image Processing* 16(8):2080–2095.

[2] Showengerdt, R. A. 2010. *Remote sensing. Models and methods for image processing*. Moscow: Technosphera. 594 p.

[3] Gonzalez, R., and R. Woods. 2012. *Digital image processing*. Moscow: Technosphera. 1104 p.

[4] Samojlin, E. A. 2014. Method of distinguishing random signals multicomponent images and impulse noise based on interchannel redundancy properties. *Digital Signal Processing* 3:2–8.

[5] Vozel, B., R. A. Kozhemiakin, V. V. Lukin, A. S. Rubel, and M. K. Tchobanou. 2015. Prediction of filtering efficiency in multichannel image processing. *17th Conference (International) on Digital Signal Processing and Its Applications Proceedings*. Moscow. 707–711.

[6] Petrov, E. P., E. V. Medvedeva, and A. P. Metelyov. 2010. Method of combined nonlinear filtration of correlated videoimages. *Nelineynyy mir* [Nonlinear World] 11:677–684.

[7] Petrov, E. P., E. V. Medvedeva, and A. P. Metelyov. 2011. Method of synthesis of video images mathematical models based on multidimensional Markov chains. *Nelineynyy mir* [Nonlinear World] 4:213–231.

[8] Petrov, E. P., I. S. Trubin, E. V. Medvedeva, and S. M. Smolskiy. 2013. Development of nonlinear filtering algorithms of digital half-tone images. *Integrated models for information communication systems and net-works: Design and development*. IGI Global. 278–304.

[9] Laletin, A. V., E. V. Medvedeva, and E. A. Ustjuzhanina. 2014. Nonlinear filtration of digital half-tone images with use of a sliding window. *Conference on Radio-Electronic Means of Receiving, Processing and Information Visualization Proceedings*. N. Novgorod. 217–221.

[10] Laletin, A. V., E. V. Medvedeva, and E. A. Ustjuzhanina. 2015. Method for improving quality of video-sequence distorted by noise. *17th Conference (International) on Digital Signal Processing and Its Applications Proceedings*. Moscow. 715–719.

Received July 9, 2015

Минимизация признакового пространства распознавания трехмерного изображения на основе стохастической геометрии и функционального анализа*

***Н. Г. Федотов*[1], *А. А. Семов*[1], *А. В. Моисеев*[2]**

fedotov@pnzgu.ru, mathematik_aleksey@mail.ru, moigus@mail.ru

[1]Пензенский государственный университет, ул. Красная, 40, г. Пенза, Россия

[2]Пензенский государственный технологический университет,
проезд Байдукова/ул. Гагарина, 1, а/11, г. Пенза, Россия

Предложен новый подход к распознаванию трехмерных (3D) изображений, основанный на современных методах стохастической геометрии и функционального анализа. Данный метод обладает рядом преимуществ, в частности позволяет описывать метрические свойства 3D объектов. Так, благодаря построению строгой математической модели аналитик может строить признаки не интуитивно, а аналитически, описывая форму объектов и их особенности (в частности, конструирование геометрических признаков). Гипертрейс-преобразование позволяет создавать инвариантное описание пространственного объекта, которое является более устойчивым к искажениям и координатным шумам, чем описание, получаемое в результате процедуры нормализации объекта. Достоверность и эффективность предлагаемого метода подтверждается как адекватно построенной математической моделью с применением современных подходов анализа и распознавания 3D изображений, так и результатами практических экспериментов, а также регистрацией разработанного программного пакета. Дано подробное описание техники сканирования гипертрейс-преобразования и его математической модели. Проанализированы основные подходы к построению и выделению информативных признаков. Предложена собственная методика минимизации признакового пространства и соответствующая ей решающая процедура. Приведены результаты практического эксперимента сравнения стохастического и детерминированного способов сканирования.

Ключевые слова: *3D распознавание образов; гипертрейс-преобразование; композиционная структура признака; минимизация признакового пространства; инвариантное описание; стохастический способ сканирования*

DOI: 10.21469/22233792.1.13.03

Feature space minimization of three-dimensional image recognition based on stochastic geometry and functional analysis*

***N. G. Fedotov*[1], *A. A. Syemov*[1], *and A. V. Moiseev*[2]**

[1]Penza State University, 40 Krasnaya st., Penza, Russia

[2]Penza State Technological University, 1-a Bajdukova proezd, Penza, Russia

Background: In recent decades, the emphasis in the analysis and pattern recognition shifts from two-dimensional to three-dimensional (3D) images, because 3D design allows to use more

*Работа выполнена при финансовой поддержке РФФИ, проект № 15-07-04484.

information about the object. Three-dimensional modeling gives possibility to see object from different angles, in particular, allows to analyze its spatial form.
Methods: A new approach to the 3D objects' recognition based on modern methods of stochastic geometry and functional analysis is proposed. This method has many advantages; in particular, it allows to describe 3D objects metric properties. Thus, due to building a rigorous mathematical model, the analyst can construct analytical and not intuitive features, describing object form and their characteristics (in particular, constructing geometric features).
Results: Hypertrace transform allows to create invariant description of spatial object, which is more resistant to distortion and coordinate noise than the description obtained as a result of the object normalization procedure. The proposed method reliability and efficiency are confirmed both an adequate constructed mathematical model by using modern approaches of 3D images analysis and recognition and practical experiments results and also the developed software package registration.
Concluding Remarks: Detailed description of hypertrace transform scan technique and its mathematical model is provided. The main approaches to construct and distinguish informative features are analyzed. Own method to minimize the feature space and its appropriate decision procedure are proposed. The practical experiment results of comparing stochastic and deterministic scan methods are presented.
Keywords: *3D image recognition; hypertrace transform; compositional structure of feature; feature space minimization; invariant description; stochastic scan method*

DOI: 10.21469/22233792.1.13.03

1 Введение

Одной из проблем распознавания образов является выработка решающей процедуры для различия одних изображений от других [1, 2]. Определение наиболее оптимального решающего правила осуществляется на основе априорной информации и обучающей выборки. При этом существуют два подхода к формированию словаря признаков. Первый подход предполагает, что алфавит классов и словарь признаков заранее предопределены и имеют четкую детерминированную структуру. Второй же подход использует методы динамического обучения словаря, наиболее активно развивавшиеся в последние 15 лет [3, 4]. Так, динамическое обучение словаря может формироваться непосредственно на обучающей выборке и может, например, значительно улучшить качество удаления шума с помощью разреженного представления изображения.

Однако предположение о полной определенности словаря признаков (первый подход) еще до распознавания изображения не вполне верно. Большие базы изображений содержат огромное количество разнообразных классов объектов, каждый из которых обладает своими собственными значимыми характеристиками и особенностями. А потому одни классы будут хорошо различаться по некоторому признаку, а другие классы — нет.

Достоинства и недостатки построения оптимальных признаков на основе априорной информации и обучающей выборке (второй подход) отмечаются, в частности, в работе [5]. Чем удачнее выбраны признаки, тем компактнее образы и успешнее решается задача обучения. Однако если признаки заранее не заданы или выбираются случайно и неудачно, то любой метод может потерять работоспособность и снизится качество распознавания. Тем не менее второй подход является более перспективным, чем первый.

Таким образом, формирование информативных числовых признаков по классам и учет их различной различающей способности (весов) для разных типов объектов являются не

менее важной и трудной задачей [6], чем разработка решающей процедуры, и могут тоже заметно повысить надежность распознавания.

Для современного этапа развития теории распознавания образов актуально расширение круга рассматриваемых задач распознавания на 3D изображения, в то время как ранее внимание исследователей было сосредоточено на решении задач анализа и распознавания двумерных (2D) изображений.

Проблема распознавания конкретно 3D изображений имеет много различных аспектов [7, 8]: проблемы анализа сцены (в том числе проблемы положения, ориентации и освещения объекта), проблемы понимания изображения, проблемы машинного зрения, а также собственно проблемы распознавания и классификации пространственных объектов на 3D изображении.

В данной статье предлагается новый подход к конструированию признаков 3D изображения на основе стохастической геометрии, дающие инвариантное описание объекта при любой его пространственной ориентации — гипертрейс-преобразование. Благодаря композиционной структуре функционалов, входящих в структуру признака, возможно построение большого числа признаков 3D изображений и применение простого решающего правила для отнесения объекта к тому или иному классу.

2 Проблема инвариантности описания трехмерного изображения

Наличие произвольной пространственной ориентации 3D изображения сильно осложняет создание его инвариантного описания. В отличие от 2D случая, поворот пространственного изображения возможен в трех плоскостях, которые являются взаимозависимыми друг от друга [9]. Так, любое вращение в 3D пространстве может быть представлено в виде композиции поворотов вокруг трех ортогональных декартовых осей, которые задаются матрицами поворота:

$$\mathbf{M}_x(\alpha) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix};$$

$$\mathbf{M}_y(\beta) = \begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix};$$

$$\mathbf{M}_z(\gamma) = \begin{pmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Так как матрицы поворота не обладают свойством коммутативности, то перемена матриц местами изменит положение объекта с одного на другое. В общем случае ориентация пространственного объекта, получаемая в результате последовательности конечных поворотов, зависит от порядка выполнения этих поворотов.

Таким образом, необходимо разработать такую схему сканирования 3D изображения, чтобы результаты его сканирования не зависели бы от пространственной ориентации анализируемого объекта. Пусть F — исходная 3D модель. Определим плоскость $B(\eta, r) = \{x | x^{\mathrm{T}}\eta = r\}$ как касательную к сфере с центром в начале координат и с радиусом r в точке (η, r), где $\eta = [\cos\varphi \cdot \sin\omega, \sin\varphi \cdot \sin\omega, \cos\omega]$ — единичный вектор в R^3, а r, ω и φ — сферические координаты.

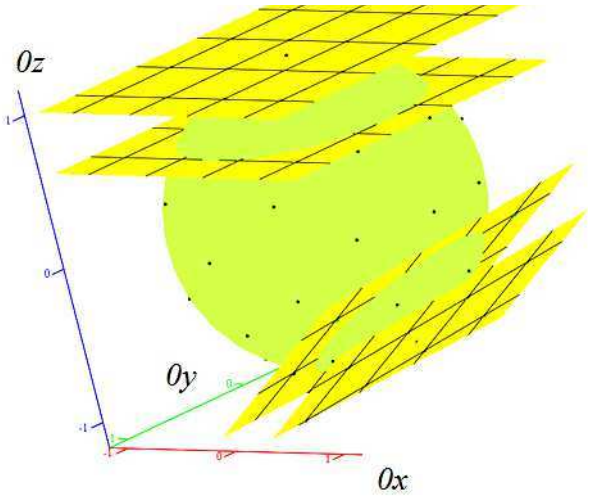


Рис. 1 Опорная сетка на сфере и соответствующие ей сетки сканирующих параллельных плоскостей

Чтобы схема сканирования 3D изображения не была привязана к форме анализируемого объекта и его пространственной ориентации, необходимо и достаточно, чтобы при произвольном 3D повороте 3D изображения получаемые плоскостями сечения не изменяли формы. Другими словами, необходимо добиться, чтобы при пространственном повороте 3D изображения все сканирующие сетки параллельных плоскостей под разными углами обзора ω и φ совпадали бы друг с другом.

Стандартный перебор всех пар углов ω и φ, которыми определяется каждая сканирующая сетка параллельных плоскостей, в топологическом смысле для непрерывного случая дает модель концентрических сфер с центром в начале координат. Каждой сканирующей сетке параллельных плоскостей на единичной сфере сопоставим точку, которая будет являться точкой касания со сферой плоскости, параллельной плоскостям данной сетки. Множество точек на сфере образуют опорную сетку (рис. 1).

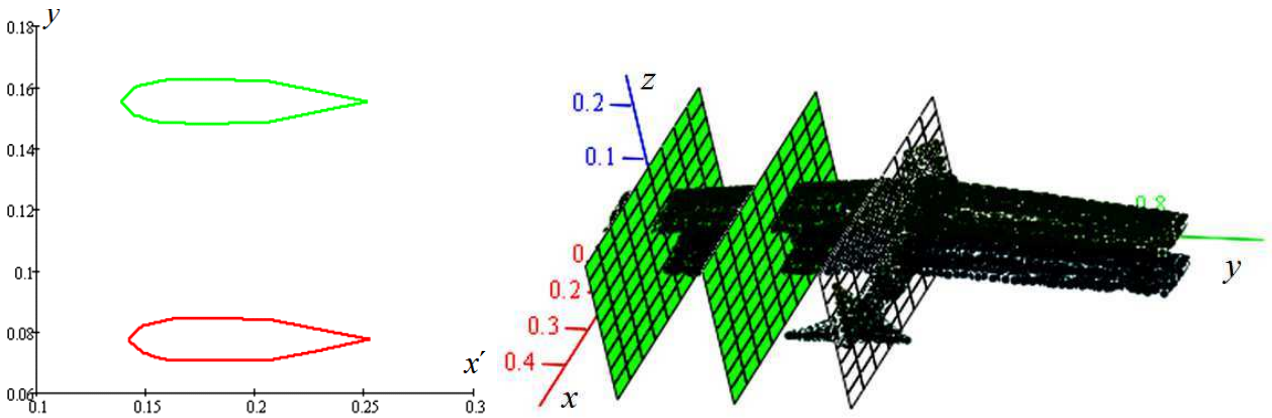
Стоит отметить, что пара углов (ω, φ) однозначно определяет узел опорной сетки, соответствующий единственной касательной плоскости и сетке сканирующих параллельных плоскостей. Если при повороте сферы вокруг своего центра опорная сетка перейдет сама в себя, то соответствующие сетки сканирующих плоскостей полностью совпадут друг с другом и получаемые сечения не изменят своей формы, поэтому вычисляемое значение признака не изменится.

Таким образом, необходимо построить опорную сетку, обладающую равномерным распределением точек на сфере, чтобы плотность плоскостей в пространстве была также равномерной для достижения наименьшей ошибки совмещении узлов опорной сетки при повороте. Равномерное распределение точек опорной сетки на сфере обеспечит отсутствие более плотных скоплений узлов на поверхности сферы в тех или иных местах, определяющих преимущественно сечения под теми или иными углами обзора. Таким образом, в общем случае при занесении результатов сканирования в матрицу (подробное описание матрицы будет дано ниже) все ее элементы будут принимать равноправное участие при вычислении значения признака 3D изображения без повышения доли влияния определенных значений элементов матрицы, так как частота появления любого элемента матрицы приблизительно одинакова (равномерный обзор 3D тела со всех сторон). Другими слова-

ми, значение вычисляемого признака не будет зависеть от ориентации 3D изображения в пространстве и будет иметь небольшое колебаний значений из-за дискретной сетки на сфере. Наибольшие отклонения будут наблюдаться тогда, когда узлы повернутой сетки будут лежать между узлами исходной сетки.

3 Математическая модель и техника сканирования гипертрейс-преобразования

Сканирование исходного пространственного объекта F осуществляется сеткой параллельных плоскостей с расстоянием Δr между плоскостями и заданными углами ω и φ (см. рис. 1). Взаимное положение 3D объекта F и каждой сканирующей плоскости $B(\eta(\omega,\varphi),r)$ характеризуется числом G по некоторому правилу HyperT : $G = \text{HyperT}\,(F \bigcap B\,(\eta(\omega,\varphi),r))$. В качестве указанной характеристики могут выступать число пересечений плоскости с исходным объектом, площадь сечения или свойства окрестности такого сечения и т. п. Другими словами, функционал HyperT характеризует свойство признака сечения [10] (рис. 2).



(*а*) Одно из сечений 3D объекта (*б*) Процесс сканирования 3D объекта плоскостями

Рис. 2 Особенности сканирования 3D объекта

Сканирование сеткой параллельных плоскостей повторяется для каждого нового значения угла обзора, определяемого выражениями $\omega + \Delta\omega$ и $\varphi + \Delta\varphi$, с тем же шагом Δr между сканирующими плоскостями. Углы ω и φ меняются согласно узлам опорной сетки.

Результат вычислений HyperT функционала зависит от трех параметров плоскости (r,ω,φ). Поэтому если каждой 2D фигуре, полученной при сечении исходной 3D модели сканирующей плоскостью, сопоставить некоторый информативный признак $\Pi\,(F_{\text{sect}})$ по правилу HyperT, то при численном анализе результат гипертрейс-преобразования удобно представить в виде 3D гипертрейс-матрицы 3ТМ, у которой ось 0φ направлена горизонтально, ось 0ω — вертикально, ось $0r$ — вглубь.

Каждая глубинная строка матрицы содержит элементы-признаки, вычисляемые по фигурам, которые получены в результате пересечения сканирующих плоскостей и исходного объекта для всех значений расстояний r при фиксированных значениях углов ω и φ. Если плоскость B не пересекает 3D изображение, т. е. $F \bigcap B\,(\eta(\omega,\varphi),r) = \varnothing$, то значение гипертрейс функционала полагают равным нулю: $\text{HyperT}\,(F \bigcap B\,(\eta(\omega,\varphi),r)) = 0$.

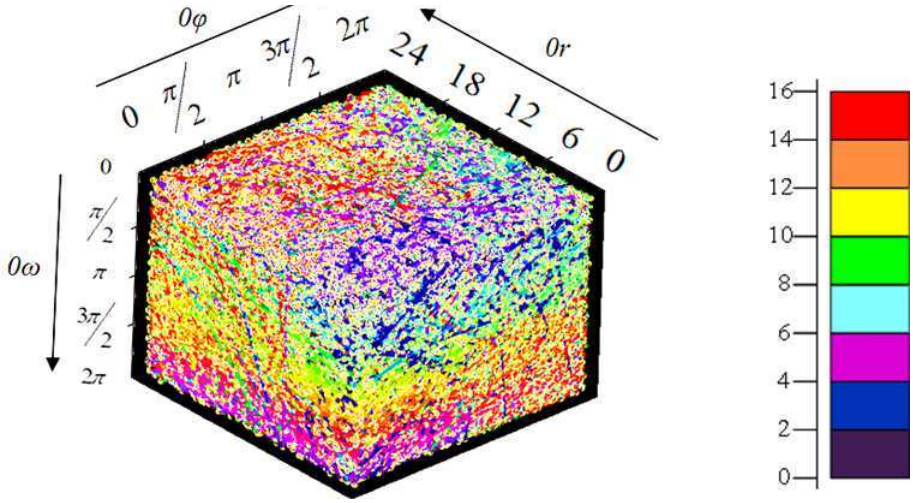


Рис. 3 Пример графического представления гипертрейс-матрицы 3TM

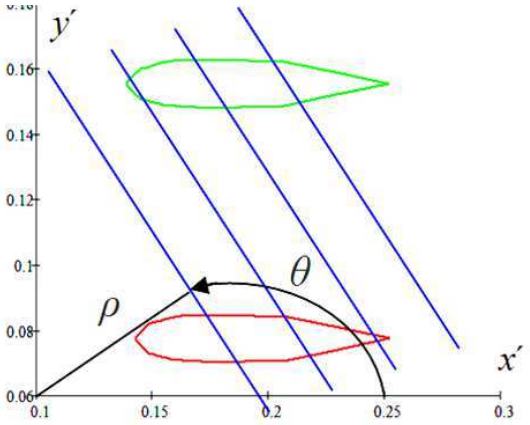
Таким образом, тройке $(\omega_i, \varphi_j, r_k)$ соответствует элемент матрицы 3TM с номером (i, j, k) и значением $\Pi(F_{\text{sect}})$, который характеризует информативный признак фигуры, полученной в сечении объекта F плоскостью $B(\eta(\omega_i, \varphi_j), r_k)$. Графическое представление гипертрейс-матрицы показано на рис. 3, где полученное в результате сканирования множество чисел G образуют точки $(\omega_i, \varphi_j, r_k)$ в системе координат с осями 0ω, 0φ и $0r$. Чем ближе к красному и теплым тонам, тем выше значение элемента матрицы; чем ближе к фиолетовому и холодным тонам, тем ниже значение данного элемента.

После заполнения 3D гипертрейс-матрицы с помощью гипердиаметрального функционала HyperP обрабатываются глубинные строки матрицы 3TM. Его можно задать, например, как $\text{HyperT} = \int G(\omega, \varphi, r)\, dr$. После обработки данная 3D гипертрейс-матрица 3TM становится двумерной матрицей 2TM, каждый столбец и строка которой представляет собой 2π-периодическую кривую. Далее применяется постолбцовая обработка матрицы 2TM посредством функционала HyperΩ, который можно задать, например, как $\text{HyperΩ} = \max_{\varphi} G(\omega, \varphi)$. В результате получается набор чисел 1TM — вектор значений, непрерывным аналогом которого будет 2π-периодическая кривая. К полученному набору чисел применяют гиперкруговой функционал HyperΘ, что приводит к появлению некоторого числа — признака изображения $\text{Res}(F)$. Этот функционал можно задать, например, амплитудой второй гармоники разложения в ряд Фурье.

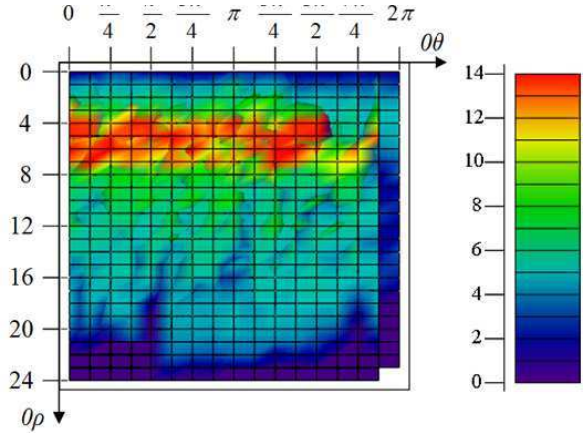
Таким образом, гипертриплетный признак 3D изображения F обладает структурой в виде композиции четырех функционалов, каждый из которых, кроме гипертрейс-функционала HyperT, при последовательном применении сокращает размерность матрицы 3TM на единицу:

$$\text{Res}(F) = \text{HyperΘ} \circ \text{HyperΩ} \circ \text{HyperP} \circ \text{HyperT}(F_{\text{sect}}).$$

Каждую 2D фигуру, получившуюся в сечении исходной 3D модели сеткой параллельных плоскостей под разными углами обзора, необходимо просканировать, чтобы извлечь какие-нибудь значимые признаки (например, периметр контура фигуры сечения и т. п.). Для нахождения признака 2D изображения сечения используется трейс-преобразование [11].



(*а*) Сканирование сечения сеткой параллельных прямых



(*б*) Пример графического представления трейс матрицы TM

Рис. 4 Процесс сканирования 2D сечения

Сканирование получаемых в сечении фигур F_{sect} осуществляется решеткой параллельных прямых $l\,(\theta, \rho)$ с расстоянием $\Delta\rho$ между линиями, где ρ и θ — полярные координаты прямой в плоскости сечения (см. рис. 2). Взаимное положение 2D изображения F_{sect} и каждой сканирующей линии $l\,(\theta, \rho)$ характеризуется числом g, вычисляемым по некоторому правилу T : $g = \mathrm{T}(F_{\text{sect}} \bigcap l(\theta, \rho))$. В качестве указанной характеристики могут выступать длина части прямой, лежащей внутри изображения, или свойства окрестности точки пересечения и т. п. Другими словами, функционал T характеризует свойство отрезков пересечений прямой 2D фигуры в плоскости сечения.

Затем сканирование производится для нового значения угла $\theta + \Delta\theta$, получившего дискретное приращение $\Delta\theta$, сеткой параллельных прямых в той же плоскости сечения F_{sect} и с тем же шагом $\Delta\rho$. К пересечению новой прямой $l(\theta + \Delta\theta, \rho)$ и сечения F_{sect} применяется то же ранее выбранное правило T. Сканирование повторяется для каждого нового угла до завершения оборота в 2π рад.

Результат вычислений трейс функционала зависит от двух параметров прямой: θ и ρ. При численном анализе результат трейс-преобразования удобно представить в виде 2D трейс матрицы TM, у которой ось 0θ направлена горизонтально, а ось 0ρ — вертикально (рис. 4). Каждый вертикальный столбец матрицы TM содержит значения, вычисляемые по всем прямым сканирующей сетки при одинаковом значении угла θ для одной и той же 2D фигуры сечения. Каждая горизонтальная строка матрицы TM содержит значения, вычисляемые для одной и той же прямой l, имеющей одинаковое расстояние до начала координат, при разных значениях угла θ в той же плоскости сечения. Если прямая l не пересекает изображение: $F_{\text{sect}} \bigcap l(\theta, \rho) = \varnothing$, то значение трейс функционала полагают равным нулю: $\mathrm{T}(F_{\text{sect}} \bigcap l(\theta, \rho)) = 0$. Таким образом, паре (θ_i, ρ_j) соответствует элемент матрицы TM с номером (i, j) и значением $\mathrm{T}(F_{\text{sect}} \bigcap l(\theta_i, \rho_j))$.

После заполнения 2D трейс матрицы с помощью диаметрального функционала P обрабатываются столбцы матрицы TM. Его можно задать, например, как $\mathrm{T} = \int g(\theta, \rho)d\rho / \max_{\rho} g(\theta, \rho)$. После этой обработки данная двумерная матрица TM становится одномерной матрицей — вектором чисел, непрерывным аналогом которого будет 2π-периодическая кривая. Затем к полученному набору чисел применяют круговой функ-

ционал Θ, который можно задать как $\Theta = \min\limits_{\theta} g(\theta)$. В результате получается некоторое число $\Pi(F_{\text{sect}})$ — признак 2D фигуры сечения F_{sect}.

Таким образом, триплетный признак 2D изображения F_{sect} обладает структурой в виде композиции трех функционалов, каждый из которых, кроме трейс функционала T, при последовательном применении сокращает размерность матрицы TM на единицу:

$$\Pi(F_{\text{sect}}) = \text{HyperT}(F_{\text{sect}}) = \Theta \circ \text{P} \circ \text{T}(F_{\text{sect}} \bigcap l(\theta, \rho)).$$

Объединяя полученные формулы для $\text{Res}(F)$ и $\Pi(F_{\text{sect}})$, окончательно получаем следующую аналитическую структуру признака 3D изображения в виде композиции некоторого множества функционалов:

$$\text{Res}(F) = \text{Hyper}\Theta \circ \text{Hyper}\Omega \circ \text{HyperP} \circ \text{HyperT}(\Theta \circ \text{P} \circ \text{T}(F_{\text{sect}} \bigcap l(\theta, \rho))).$$

Благодаря композиционной структуре функционалов, входящих в структуру $\Pi(F_{\text{sect}})$ и $\text{Res}(F)$, возможно получение огромного числа признаков, причем возможно конструирование признаков, описывающих те или иные геометрические характеристики 3D объекта, что облегчает задачу анализа свойств 3D изображений и построения информативных признаков.

Стоит отметить, что расположение системы координат в плоскости сечения и ее ориентация относительно фигуры сечения совершенно неважны, так как трейс-преобразование полностью инвариантно к группе движений и масштабированию 2D изображения [12].

Таким образом, данный метод обладает определенной универсальностью, так как схема сканирования не привязана к геометрическим особенностям исходной модели, а благодаря большому числу используемых видов функционалов и их композиционной структуре можно подбирать и конструировать различные признаки, которые будут наиболее эффективны при распознавании заданной базы объектов. Предлагаемая методика ориентирована на объекты любой сложности и конфигурации.

4 Минимизация размерности признакового пространства

Как отмечалось выше, подход на основе стохастической геометрии позволяет автоматически генерировать большое количество гипертриплетных признаков, отображающих как геометрические, так и абстрактные характеристики 3D изображения. Однако сформированная таким образом система гипертриплетных признаков, как правило, избыточна. Одни признаки имеют высокую различающую силу, тогда как другие — нет. При этом для одних классов эффективны одни признаки, а для других классов — другие. Кроме того, многие признаки могут коррелировать друг с другом, тем самым снижая общую эффективность распознавания.

Также стоит отметить, что создание излишнего большого количества признаков не только не повышает эффективность распознавания, но и снижает скорость работы распознающей системы.

Таким образом, целесообразно разработать алгоритм, который после генерации признаков минимизирует размерность их пространства для выделения наиболее информативных из них.

Можно выделить два основных подхода к построению эффективного множества признаков изображений [13, 14]. Первый подход заключается в том, чтобы строить заранее

известное малое количество признаков, обладающих большой информативностью. С позиции второго подхода из большого числа построенных признаков по некоторому правилу отбирается как можно меньшее количество наиболее информативных признаков.

Минусом первого подхода является отсутствие единой логической системы, так как такие методы основаны, как правило, на эвристике и эмпирике разработчика распознающего алгоритма, поэтому на практике трудно выявить малое количество информативных признаков, которые будут эффективно распознавать 3D изображения для большинства практических задач, так как геометрия реальных 3D объектов весьма обширна и сложна.

Касательно второго подхода в настоящее время разработано множество различных критериев отбора эффективных и значимых признаков, основанных на методах математической статистики и информатики. Данный подход является более гибким и универсальным, так как в каждом конкретном случае информативность признаков оценивается исходя из представленной базы 3D объектов.

В данном разделе будет описан алгоритм, согласно которому можно сократить большое количество признаков до нескольких наиболее информативных (в русле идей второго подхода). Также благодаря композиционной структуре признака аналитик может заранее создавать определенные признаки, которые с высокой вероятностью будут являться информативными (в русле идей первого подхода), такие как площадь поверхности 3D объекта, его объем, максимальная площадь и периметр сечения, радиус описанной сферы и т. п.

Количественной мерой для определения информативности отдельного признака Res_i может служить количество информации Info(Res), извлекаемой при распознавании 3D изображения. Она равна разности между энтропией $H(F)$ распределений плотности вероятности образов F и усредненной по всем изображениям неопределенностью решения, которая определяется полной условной энтропией образов F_i.

Однако в рассматриваемом случае оценка информативности признака данным способом невозможна ввиду огромного объема вычислений. Так, система способна автоматически генерировать $64^6 = 68\,719\,476\,736$ различных признаков при использовании 64 различных видов функций для каждого из 6 функционалов композиционной структуры признака [15]. Очевидно, что задача определения даже небольшого числа информативных из всей совокупности признаков не разрешима за реальное время в рамках определения количественной меры энтропии. Кроме того, не всегда возможно получить численные значения вероятностей, необходимых для определения $H(F)$ и $\mathrm{Info}(x)$.

Также стоит отметить, что концепция минимальной энтропии основывается на предположении о нормальности распределения образов, составляющих заданные классы, что далеко не всегда верно. Кроме того, в задаче классификации 3D изображений законы распределений плотности вероятности образов не известны, так как базы данных формируются исходя из контекста решаемой задачи, определяемого конкретными условиями деятельности того или иного субъекта.

Поэтому целесообразно использовать подходы, для которых не нужно знать плотность распределения вероятности 3D изображений. Будем определять информативность того или иного признака исходя из данных обучающей выборки 3D объектов.

В основе данного предположения лежит гипотеза компактности: 3D изображения одного и того же класса в признаковом пространстве обычно располагаются в геометрически близкие точки, образуя «компактные» сгустки. Другими словами, схожие объекты гораздо чаще лежат в одном классе, чем в разных, и при этом обладают свойством хорошей отделимости:

(1) множества разных образов соприкасаются в сравнительно небольшом числе точек (или вообще не соприкасаются);
(2) существуют точки в признаковом пространстве, которые не будут принадлежать ни к одному из классов (или равновероятно принадлежать обоим классам);
(3) границы классов имеют сравнительно плавную форму без глубоких выступов в пределы других классов.

Таким образом, используемый в данной работе алгоритм минимизации размерности признакового пространства был разработан исходя из логики сравнения 3D объектов, учитывая все вышесказанное. Его суть заключается в следующем.

Рассмотрим множество $M = \bigcup_{i=1}^{m} C_i$, состоящее из m подмножеств (классов) C_i, при этом в подмножестве C_i содержится h_i количество элементов (изображений). Выберем из данного множества подмножество $M' \subset M$, мощность которого равна $\sum_{i=1}^{m} h_i/2$ для обучения системы, т. е. $M' = \bigcup_{i=1}^{m} A_i$. Оставшиеся подмножества B_i будут нужны для испытания обученной системы, контроля ее качества.

Обозначим через $\mathrm{Res}_k^{A_i(s)}$ гипертриплетный признак k-го вида, вычисленный для s-го представителя i-го класса A_i. Среднее значение k-го вида признака для всех изображений множества A_i равно:

$$\mu_k^{A_i} = \frac{2}{h_i} \sum_{s=1}^{h_i} \mathrm{Res}_k^{A_i(s)}.$$

Среднеквадратическое отклонение k-го признака по множеству A_i равно:

$$\sigma_k^{A_i} = \sqrt{\frac{2}{h_i} \left(\sum_{s=1}^{h_i} \left(\mathrm{Res}_k^{A_i(s)} - \mu_k^{A_i} \right)^2 \right)}.$$

Информативными признаками будут те, которые позволяют различать как можно больше классов 3D объектов между собой. Другими словами, среднее значение признака для одного класса будет как можно более удалено от среднего значения того же признака для любого другого класса. При этом количество совпадений представителей разных классов должно быть как можно меньше.

В связи с вышеизложенным были разработаны две процедуры: отбор потенциально эффективных признаков по количеству как можно меньшего совпадения представителей разных классов между собой и выделение информативных признаков по удаленности друг от друга их средних значений по классам.

Для отбора потенциально эффективных признаков необходимо внутри каждого класса изображений произвести расчет их усредненных значений, а также статистику расчета их среднеквадратического колебания отдельно по каждому признаку. Для этого рассчитывается показатель $p(A_i, k)$, который определяет меру неподобия k-го признака для i-го класса A_i. Он состоит из двух частей:

$$p(A_i, k) = \frac{q_1(A_i, k) + q_2(A_i, k)}{h_i}.$$

Первая часть коэффициента учитывает количество 3D изображений, признаки $\mathrm{Res}_k^{A_i(s)}$ которых не будут попадать в соответствующие диапазоны колебания среднего значения

k-го признака для своего i-го класса A_i: $q_1(A_i(s),k)$ увеличивается на 1, если $\mathrm{Res}_k^{A_j(s)} < \mu_k^{A_i} - \sigma_k^{A_i} \vee \mu_k^{A_i} + \sigma_k^{A_i} < \mathrm{Res}_k^{A_j(s)}$ для $i = j$.

Показатель $q_1(A_i(s),k)$ показывает количество 3D объектов s, которые будут неправильно классифицированы, так как слишком далеко отстоят от среднего представителя своего класса.

Вторая часть коэффициента учитывает количество 3D изображений, признаки $\mathrm{Res}_k^{A_j(s)}$ которых попадут в соответствующие диапазоны колебания среднего значения k-го признака для другого i-го класса A_i: $q_2(A_i(s),k)$ увеличивается на 1, если $\mu_k^{A_i} - \sigma_k^{A_i} \leqslant \mathrm{Res}_k^{A_j(s)} \leqslant \mu_k^{A_i} + \sigma_k^{A_i}$ для $i \neq j$.

Показатель $q_2(A_i(s),k)$ показывает количество 3D объектов s, которые будут неправильно классифицированы, так как слишком близко находятся к среднему представителю другого класса.

Коэффициент p представляет собой матрицу весов, рассчитанную для каждого типа признака k ($k = 1,\ldots,\mathrm{col}$) в зависимости от класса объектов A_i ($i = 1,\ldots,m$). Данная матрица показывает различающую силу признаков с точки зрения уровня потенциальных ошибок в классификации объектов.

Стоит отметить, что коэффициент $p(A_i,k)$ всегда будет лежать в единичном отрезке: $0 \leqslant p(A_i,k) \leqslant 1$. Поэтому данный показатель может рассматриваться как вероятность того, насколько этот признак потенциально неинформативен для данного класса. Чем выше коэффициент неподобия $p(A_i,k)$, тем меньшей различающей силой обладает k-й признак для i-го класса A_i. В связи с этим целесообразно задать некоторый порог δ, чтобы из всей совокупности признаков выделить потенциально эффективные, отсеяв заведомо неинформативные признаки:

$$\frac{\sum_{i=1}^{m} p(A_i,k)}{m} \leqslant \delta.$$

Таким образом, данная процедура позволяет не только произвести отбор потенциально эффективных признаков, но и указать их различающую силу по классам.

Далее для учета корреляции признаков друг с другом нужно произвести расчет элементов матрицы парной корреляции по полученной совокупности средних значений каждого признака по классам. Выделяются те признаки, которые имеют значение коэффициента парной корреляции не ниже 0,7. Среди выделенных пар множества признаков удаляются те, которые имеют большее значение уже подсчитанной суммы $\sum_{i=1}^{m} p(A_i,k)$.

Дальнейший отбор информативных признаков производится из полученной сокращенной совокупности признаков с учетом распределения их средних значений по классам на всем интервале, чтобы в целом все признаки были удалены как можно дальше друг от друга. Так как колебание дисперсии признаков уже учтено по классам (в матрице весов выделены только наиболее информативные с точки зрения как можно меньшего пересечения границ классов между собой), то далее достаточно для каждого признака произвести сортировку только их усредненных значений по классам:

$$\mathrm{sort}\left(\mu_k^{A_1}, \mu_k^{A_2}, \ldots, \mu_k^{A_m}\right) \to \left(\xi_k^{A_1}, \xi_k^{A_2}, \ldots, \xi_k^{A_m}\right),$$

где $\xi_k^{A_i}$ — среднее значение k-го признака по i-му классу A_i, отсортированное по возрастанию ($\xi_k^{A_1} \leqslant \xi_k^{A_2} \leqslant \cdots \leqslant \xi_k^{A_m}$).

Далее находится разница между значениями соседних элементов:

$$\Delta\left(\xi_k^{A_1}, \xi_k^{A_2}, \ldots, \xi_k^{A_m}\right) \to \left(\xi_k^{A_2} - \xi_k^{A_1}, \xi_k^{A_3} - \xi_k^{A_2}, \ldots, \xi_k^{A_m} - \xi_k^{A_{m-1}}\right).$$

Чем меньше значение элемента $\Delta\xi_k^i = \xi_k^{A_i} - \xi_k^{A_{i-1}}$ $(i = 2, \ldots, m)$ и тем больше количество таких элементов, тем хуже различающая сила признака. Чтобы не производить лишний раз сортировку и при этом оценить уровень значения нескольких наиболее худших представителей, достаточно рассчитать, например, нижнюю границу интервала колебания среднего значения признаков $\Delta\xi_k^i$ $(i = 2, \ldots, m)$:

$$\text{border}(k) = \text{mean}\,(\Delta\xi_k) - \text{stdev}\,(\Delta\xi_k)\,,$$

где $\text{mean}\,(\Delta\xi_k) = (2/(m-1))\sum_{i=1}^{m-1}\Delta\xi_k^i = 2\left(\xi_k^{A_m} - \xi_k^{A_1}\right)/(m-1)$ — среднеарифметическое чисел $\Delta\xi_k^i$, а $\text{stdev}\,(\Delta\xi_k) = \sqrt{(2/(m-1))\left(\sum_{i=1}^{m-1}\left(\Delta\xi_k^i - \text{mean}\,(\Delta\xi_k)\right)^2\right)}$ — среднеквадратическое чисел $\Delta\xi_k^i$.

Критерий отбора признаков следующий: чем выше значение $\text{border}(k)$, тем выше различающая сила k-го признака и его информативность. Поэтому программа отбирает z лучших представителей по данному критерию, где порог z признаков задает аналитик машине еще до начала работы исходя из количества классов изображений в базе данных, а также требований к точности результатов и времени их получения.

Вторая предложенная процедура выделяет из всех потенциально информативных некоррелируемых признаков только те, которые дают как можно менее близкий друг ко другу набор средних значений признаков по классам. В этом случае учитывается различающая сила признаков с точки зрения их информативности.

Построенный согласно указанным двум процедурам алгоритм сокращения размерности признакового пространства позволяет получать набор информативных признаков с указанием для каждого из них его различающей силы с точки зрения потенциального уровня ошибок (значение весового коэффициента неподобия). При необходимости данный набор можно сводить к минимуму, указывая минимальное количество z требуемых признаков, отобранных с точки зрения их потенциальной эффективности.

Вторая предложенная процедура эффективно дополняет первую процедуру, так как они преследуют прямо противоположные цели: первая отбирает информативные признаки с точки зрения уровня их потенциальной неэффективности, тогда как вторая стремится выделять признаки с точки зрения их потенциальной эффективности. В тех случаях, где плохо сработает первая процедура (например, определенные теоретические примеры конструкций), вторая должна дать хорошие результаты, и наоборот.

5 Решающая процедура

«Сходство» 3D изображений между собой определяется функцией расстояния $\rho(\text{desk}(x), \text{desk}(x'))$ между двумя векторами дескрипторов признаков образов $\text{desk}(x)$ в пространстве объектов X. При выполнении гипотезы компактности класс 3D изображения может быть также определен как класс усредненного изображения множества сходных видов пространственных объектов, являющегося наиболее близким к исходному изображению в смысле расстояния $\rho(x, x')$.

Решающая процедура построена таким образом, что может как учитывать, так и не учитывать весовые коэффициенты для каждого информативного гипертриплетного признака. Ее суть состоит в следующем.

Обозначим через t тестовое 3D изображение из какого-либо подмножества B_i. Тогда его k-й признак будет равен Res_k^t. Расстояние между тестовым 3D изображением и i-м

классом (множеством A_i) с учетом весовых значений определяется следующим образом:

$$d(t, A_i) = \sum_k p(A_i, k) \frac{|\mathrm{Res}_k^t - \mu_k^{A_i}|}{\sigma_k^{A_i}}.$$

Без учета весовых значений данная формула будет выглядеть так:

$$d(t, A_i) = \sum_k \frac{|\mathrm{Res}_k^t - \mu_k^{A_i}|}{\sigma_k^{A_i}}.$$

Распознающая система тестовое изображение t относит к классу A_j, если

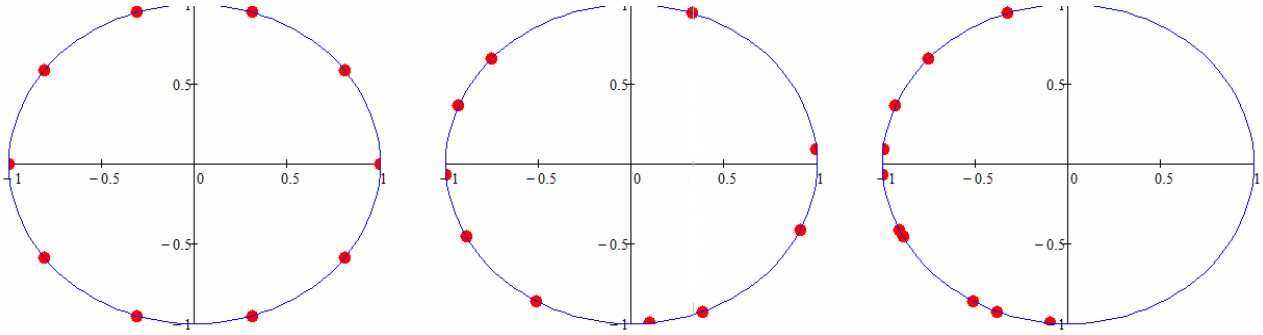
$$d(t, A_j) = \min_i d(t, A_i).$$

Таким образом, еще один существенный плюс в пользу гипертриплетных признаков заключается в том, что при опоре на большое их количество применяются простые решающие правила для распознавания 3D изображений. При этом, что немаловажно, при определении принадлежности тестового изображения учитываются весовые коэффициенты для каждого информативного признака в зависимости от класса 3D объекта.

6 Преимущества стохастического способа сканирования трехмерных изображений

При анализе и распознавании 2D изображений сканирование со случайными параметрами улучшает соотношение «надежность–быстродействие» по сравнению с фиксированной разверткой [12]. Аналогичное свойство справедливо также и при анализе и распознавании 3D изображений. Покажем преимущество стохастического способа анализа перед детерминированными развертками на простом примере бросания точки на окружность, который очень хорошо раскроет общую идею.

Задача эксперимента состоит в том, чтобы оценить минимальную погрешность отклонения случайно бросаемой точки на окружность от множества расставленных на ней точек. Для фиксированной развертки из N точек (рис. 5, $а$) на окружности образуются N дуг равной длины $2\pi/N$. Поэтому минимальное отклонение в худшем случае составит π/N, когда бросаемая точка будет лежать в центре дуги между двумя точками исходной развертки.



($а$) ($б$) ($в$)

Рис. 5 Расстановка точек на окружности стохастическим и детерминированным способом

Для стохастического способа расстановки точек дело обстоит иначе (рис. $5, б$). Не уменьшая общности, пусть бросаемая точка попала в место $(0; 1)$. Максимальная погрешность равна π, когда бросаемая точка и наиболее удаленная точка развертки будут являться диаметрально противоположными. Так как точки левой и правой половины окружности равноудалены от точки $(0; 1)$ и дают одинаковое симметричное отклонение в пределах до π, то анализ минимального отклонения бросаемой точки от множества построенных точек на окружности равносилен анализу минимального отклонения для точек только для одной половины окружности, например, если спроецировать горизонтально точки правой половины окружности на левую (рис. $5, в$). В этом случае для равномерного распределения ожидаемое минимальное отклонение для N точек будет равно $\pi/(N+1)$, так как они между точкой $(0; 1)$ и диаметрально удаленной $(0; -1)$ образуют $N+1$ дугу с ожидаемой средней длиной $\pi/(N+1)$.

Данные рассуждения аналогичны и справедливы для любой бросаемой точки на окружности, отличной от $(0; 1)$. Так как $\pi/(N+1) < \pi/N$, то стохастический способ расстановки точек более эффективен, чем детерминированный. Проведенные практические эксперименты в среде пакета MathCAD 15 M030 показали справедливость приведенных выше теоретических оценок. Другие аналогичные теоретические оценки, подтверждающие эффективность и преимущество методов стохастической геометрии, можно найти в [16].

Ниже приведены результаты эксперимента для доказательства преимуществ стохастического способа сканирования перед детерминированной разверткой с точки зрения соотношения «точность–быстродействие». В качестве 3D изображения был взят объект «паук» под номером m_{16} из известной базы данных Принстонского университета *The Princeton Shape Benchmark* [17]. В качестве критерия оценки результатов работы алгоритма был использован коэффициент удельной погрешности вычисления признака:

$$\beta = \frac{|\mathrm{TrueFeat} - \mathrm{CalcFeat}|}{\mathrm{TrueFeat}},$$

где TrueFeat — истинное значение признака, а CalcFeat — вычисленное значение признака.

В качестве анализируемого признака была выбрана максимальная длина отрезка, который может быть помещен вовнутрь 3D объекта. Точное значение данного признака можно вычислить для произвольно взятого 3D изображения, перебрав все попарные расстояния между вершинами 3D модели. Описание данного признака приведено ниже:

$$\mathrm{Res}(F) = \mathrm{Hyper}\Theta \circ \mathrm{Hyper}\Omega \circ \mathrm{HyperP} \circ \mathrm{HyperT}(\Theta \circ \mathrm{P} \circ \mathrm{T}).$$

Здесь

$$\mathrm{T}(F_{\mathrm{sect}} \bigcap l(\theta, \rho)) = \max_t(f(\theta, \rho, t));\ \mathrm{P} = \max_\rho g(\theta, \rho);\ \Theta = \max_\theta g(\theta);$$

$$\mathrm{HyperT}(F \bigcap B(\eta(\omega, \varphi), r)) = \Pi(F_{\mathrm{sect}}) = G(\omega, \varphi, r);\ \mathrm{HyperP} = \mathrm{Row3D} \cdot \Delta r;$$

$$\mathrm{Hyper}\Omega = \max_\omega G(\omega, \varphi);\ \mathrm{Hyper}\Theta = \max_\varphi G(\varphi),$$

где $f(\theta, \rho, t)$ — длина t-го отрезка, высекаемого ρ-й прямой под θ-м углом наклона в плоскости сечения F_{sect}; $\Pi(F_{\mathrm{sect}}) = G(\omega, \varphi, r)$ — признак сечения (максимальная длина отрезка), получаемого пересечением r-й плоскости $B(\eta(\omega, \varphi), r)$ под парой (ω, φ) углов обзора; Row3D — количество ненулевых элементов глубинных строках матрицы 3ТМ (ось $0r$).

Эксперимент состоял из двух блоков. Первый блок оценивал при одинаковых параметрах сканирования, на сколько процентов увеличится точность распознавания объекта (точность вычисления признака) при использовании стохастического сканирования по

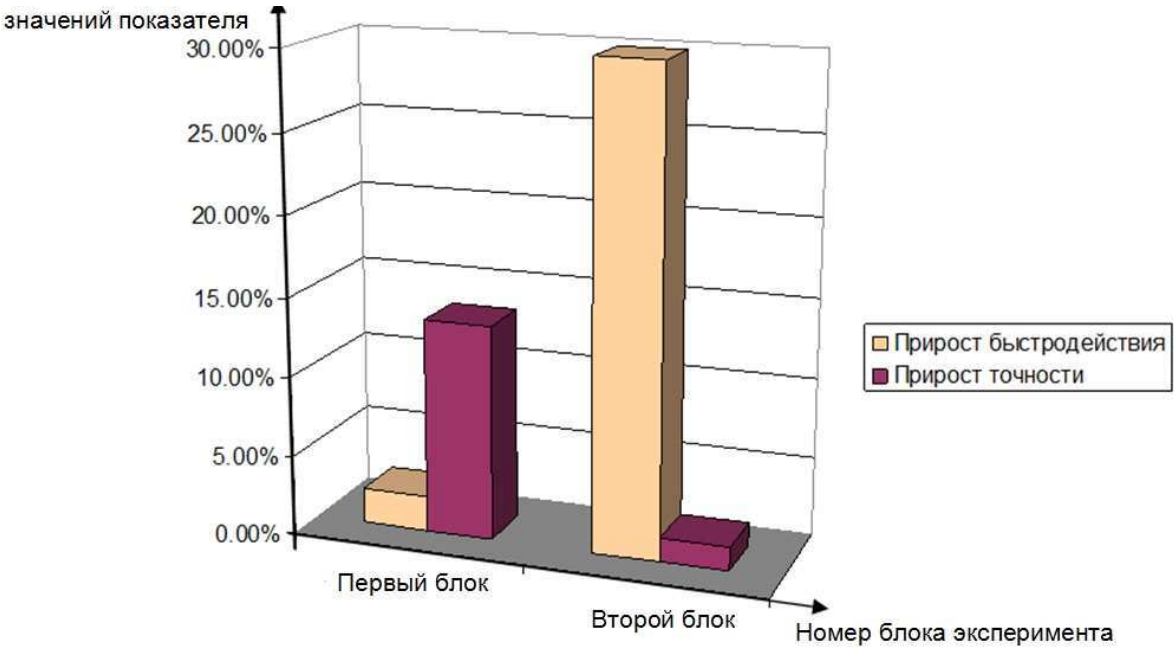


Рис. 6 Сравнительная диаграмма результатов отношений стохастического сканирования и детерминированной развертки

сравнению с детерминированной разверткой. Второй блок состоял в определении таких параметров сканирования, чтобы за меньшее время работы алгоритма (шаг сканирования крупнее) достичь приблизительно тот же уровень точности для стохастического сканирования, как и для детерминированной развертки. Горизонтальная ось показывает результаты экспериментов каждого блока. Положительные значения вертикальной оси показывает преимущество стохастического способа сканирования перед детерминированным, отрицательные значения — наоборот:

$$\Delta_{i,j} = \frac{S_{i,j} - D_{i,j}}{S_{i,j}},$$

где $S_{i,j}$ — значение показателя, вычисленного стохастическим способом сканирования 3D изображения; $D_{i,j}$ — значение показателя, вычисленного детерминированным способом сканирования 3D изображения. При $i = 1$ показатель определяет время вычисления признака, при $i = 2$ — значение признака. Переменная j определяет блок эксперимента.

Как видно из представленной диаграммы (рис. 6), преимущество стохастического сканирования перед детерминированной разверткой очевидно. При этом во втором блоке эксперимента прирост вычисляемого показателя (прирост быстродействия) заметно выше, чем в первом блоке (точность вычислений признака). При этом при упоре на быстродействие прирост эффективности заметно выше, чем при упоре на точность вычислений признака. Это объясняется комбинаторным сокращением числа сканирований из-за композиционной структуры признака. Стоит отметить также, что возможность регулирования такого свойства гипертрейс-преобразования, как выбор между скоростью вычислений и точностью, повышает гибкость стохастического распознавания 3D изображений. Другие эксперименты по проверке свойств гипертрейс-преобразования при анализе и распознавании 3D изображений можно найти в работах [18, 19].

7 Заключение

В настоящей статье для решения задачи анализа и распознавания 3D изображений впервые был предложен новый подход с позиции стохастической геометрии и функционального анализа, который позволяет анализировать пространственные объекты без предварительного их упрощения и построения проекций на плоскости, анализируя непосредственно их 3D форму. Данный подход обладает определенной универсальностью, так как техника сканирования плоскостями не привязана к геометрическим особенностям 3D изображения, поэтому гипертрейс-преобразование способно эффективно распознавать 3D объекты любой формы и структуры.

Новое геометрическое гипертрейс-преобразование благодаря описанному методу построения гипертриплетных признаков позволяет не только создавать инвариантное описание 3D изображения, но и анализировать его особенности и геометрию поверхности.

Разработанная процедура минимизации признакового пространства позволяет не только отбирать заданное число информативных признаков, но и присваивать каждому из них весовой коэффициент, обозначающий его различающую силу в зависимости от предъявляемого класса 3D изображения. Также стоит отметить, что цель статьи состояла не в разработке универсальной процедуры минимизации признакового пространства, а в возможности применения данной процедуры при стохастическом методе распознавания, которая бы обладала дополнительным преимуществом: имелась возможность учитывать весовые значения коэффициентов признаков в зависимости от классов 3D изображений. Наличие данного свойства позволяет повысить надежность стохастического распознавания 3D изображений гипертрейс-преобразованием.

Используя особенности конструирования гипертриплетных признаков, можно использовать простую процедуру минимизации признакового пространства и решающее правило. Опора на минимальный набор эффективных признаков значительно сокращает время работы распознающего алгоритма

Ввиду сложности разработки в общем виде единой методологии отбора информативных некоррелируемых признаков и ее оценки, а также ограниченного объема статьи, в данной работе отсутствует сравнение процедуры минимизации признакового пространства с другими подходами.

Авторы планируют развить данный метод для анализа не только бинарных (контурных) и монохромных 3D изображений [15], но также и цветных и текстурных 3D изображений. Аналогичные результаты уже были получены при анализе цветных и текстурных 2D изображений в [20–22]. Интеллектуальный уровень гипертрейс-преобразования может быть повышен благодаря развитию метода для интеллектуального анализа и распознавания деформированных и поврежденных 3D объектов, для изучения топологии их поверхности. Уже есть первая публикация в данном направлении [23].

Литература

[1] *Ferreira M. R. P., de Carvalho F. de A. T.* Kernel-based hard clustering methods in the feature space with automatic variable weighting // Original Research Article Pattern Recognition, 2014. Vol. 47. Iss. 9. P. 3082–3095.

[2] *Vasil'ev K. K., Dement'ev V. E., Andriyanov N. A.* Doubly stochastic models of images // Pattern Recognition Image Anal. Adv. Math. Theory Appl., 2015. Vol. 25. No. 1. P. 105–110.

[3] *Mairal J., Bach F., Ponce J., Sapiro G.* Online dictionary learning for sparse coding // 26th Annual Conference (International) on Machine Learning (ICML) Proceedings, 2009. P. 689–696.

[4] *Rubinstein R., Peleg T., Elad M.* Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model // IEEE Trans. Signal Processing, 2013. Vol. 61, No. 3. P. 661–677.

[5] *Xie L., Li D., Simske S. J.* Feature dimensionality reduction for example-based image super-resolution // J. Pattern Recognition Res., 2011. Vol. 6. No. 2. P. 130–139.

[6] *Yildiz O. T.* On the feature extraction in discrete space // Original Research Article Pattern Recognition, 2014. Vol. 47. Iss. 5. P. 1988–1993.

[7] *Zhang Y., Song S., Tan P., Xiao J.* PanoContext: A whole-room 3D context model for panoramic scene understanding // 13th European Conference on Computer Vision (ECCV 2014) Proceedings. Zurich, Switzerland, 2014. Part IV. Vol. 8694. P. 668–686.

[8] *Федотов Н. Г., Семов А. А., Курносов А. А.* Проблемы распознавания 3D изображений у машин и людей: сравнительная характеристика // Проблемы информатики в образовании, управлении, экономике и технике: Сб. статей XIV Междунар. науч.-технич. конф. — Пенза: Изд-во АНИОО «Приволжский дом знаний», 2014. С. 185–193.

[9] *Liu K., Skibbe H., Schmidt T., Blein T., Palme K., Brox T., Ronneberger O.* Rotation-invariant HOG descriptors using fourier analysis in polar and spherical coordinates // Int. J. Computer Vision, 2014. Vol. 106. Iss. 3. P. 342–364.

[10] *Fedotov N. G., Ryndina S. V., Syemov A. A.* Trace transform of spatial images // 11th Conference (International) on Pattern Recognition and Image Analasis: New Information Technologies (PRIA-11) Proceedings. — Samara: IPSI RAS, 2013. Vol. I. P. 186–189.

[11] *Fedotov N. G.* The theory of image-recognition features based on stochastic geometry // Pattern Recognition Image Anal. Adv. Math. Theory Appl., 1998. Vol. 8. No. 2. P. 264–266.

[12] *Федотов Н. Г.* Теория признаков распознавания образов на основе стохастической геометрии и функционального анализа. — М.: Физматлит, 2009. 304 с.

[13] *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989. 607 с.

[14] *Hastie, T., Tibshirani R., Friedman J.* The elements of statistical learning: Data mining, inference and prediction. — 2nd ed. — Springer-Verlag, 2009. 746 p.

[15] *Федотов Н. Г., Семов А. А.* Программный комплекс анализа и распознавания 3D изображений на основе пространственного трейс-преобразования со случайными параметрами сканирования. Свидетельство об официальной регистрации программ для ЭВМ № 2015612257 Роспатента от 16.02.15.

[16] *Santalo L. A.* Integral geometry and geometric probability. — 2nd ed. — New York, NY, USA: Cambridge University Press, 2004. 428 p.

[17] Princeton Shape Benchmark of 3D models database. http://shape.cs.princeton.edu/benchmark/.

[18] *Семов А. А.* Экспериментальная проверка свойств 3D трейс-преобразования // XXI век: итоги прошлого и проблемы настоящего плюс. Научно-методический журнал. Серия: технические науки. Информационные технологии, 2014. Вып. 03(19). С. 83–89.

[19] *Fedotov N. G., Ryndina S. V., Semov A. A.* Trace transform of three-dimensional objects: Recognition, analysis and database search // Pattern Recognition Image Anal. Adv. Math. Theory Appl., 2014. Vol. 24. No. 4. P. 566–574.

[20] *Fedotov N. G., Mokshanina D. A.* Recognition of halftone textures from the standpoint of stochastic geometry and functional analysis // Pattern Recognition Image Anal. Adv. Math. Theoryd Appl., 2010. Vol. 20. No. 4. P. 551–556.

[21] *Fedotov N. G., Mokshanina D. A.* Recognition of images with complex half-tone texture // Measurement Techniques, 2011. Vol. 53. No. 11. P. 1226–1232.

[22] *Fedotov N., Romanov S., Goldueva D.* Application of triple features theory to the analysis of halftone images and colored textures. Feature construction along stochastic geometry and functional analysis // Computer Inform. Sci., 2013. Vol. 6. No. 4. P. 17–24.

[23] *Федотов Н. Г., Семов А. А., Моисеев А. В.* Интеллектуальные возможности гипертрейс-преобразования: конструирование признаков с заданными свойствами // Машинное обучение и анализ данных, 2014. Т. 1. № 9. С. 1200–1214.

Поступила в редакцию 15.06.2015

References

[1] Ferreira, M. R. P., and F. de A. T. de Carvalho. 2014. eKernel-based hard clustering methods in the feature space with automatic variable weighting. *Original Research Article Pattern Recognition* 47(9):3082–3095.

[2] Vasil'ev, K. K., V. E. Dement'ev, and N. A. Andriyanov. 2015. Doubly stochastic models of images. *Pattern Recognition Image Anal. Adv. Math. Theory Appl.* 25(1):105–110.

[3] Mairal, J., F. Bach, J. Ponce, and G. Sapiro. 2009. Online dictionary learning for sparse coding. *26th Annual Conference (International) on Machine Learning (ICML) Proceedings*. 689–696.

[4] Rubinstein, R., T. Peleg, and M. Elad. 2013. Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model. *IEEE Trans. Signal Processing* 61(3):661–677.

[5] Xie, L., D. Li, S. J. Simske. 2011. Feature dimensionality reduction for example-based image super-resolution. *J. Pattern Recognition Res.* 6(2):130–139.

[6] Yildiz, O. T. 2014. On the feature extraction in discrete space. *Original Research Article Pattern Recognition* 47(5):1988–1993.

[7] Zhang, Y., S. Song, P. Tan, and J. Xiao. 2014. PanoContext: A whole-room 3D context model for panoramic scene understanding *13th European Conference on Computer Vision (ECCV 2014) Proceedings*. Zurich, Switzerland. 8694(IV):668–686.

[8] Fedotov, N. G., A. A. Syemov, and A. A. Kurnosov. 2014. Three-dimensional images recognition problems at machines and peoples: Comparative characteristic. *14th Scientific and Technical Conference (International) on Problems of Informatics in Education, Management, Economics and Technology Proceedings*. Penza: Privolzskiy Dom Znaniy Publs. 185–193.

[9] Liu, K., H. Skibbe, T. Schmidt, T. Blein, K. Palme, T. Brox, and O. Ronneberger. 2014. Rotation-invariant HOG descriptors using fourier analysis in polar and spherical coordinates. *Int. J. Computer Vision* 106(3):342–364.

[10] Fedotov, N. G., S. V. Ryndina, and A. A. Syemov. 2013. Trace transform of spatial images. *11th Conference (International) on Pattern Recognition and Image Analasis: New Information Technologies Proceedings*. Samara: IPSI RAS. I:186–189.

[11] Fedotov, N. G. 1998. The theory of image-recognition features based on stochastic geometry. *Pattern Recognition Image Anal. Adv. Math. Theory Appl.* 8(2):264–266.

[12] Fedotov, N. G. 2009. *The theory of patterns recognition features based on stochastic geometry and functional analysis*. Moscow: Fizmatlit. 304 p.

[13] Ayvazyan, S. A., V. M. Bukhshtaber, I. S. Enyukov, and L. D. Meshalkin. 1989. *Applied statistics: Classification and dimension reduction*. Moscow: Finance and Statistics. 607 p.

[14] Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: Data mining, inference and prediction*. 2nd ed. Springer-Verlag. 746 p.

[15] Fedotov, N. G., and A. A. Syemov. February 16, 2015. Software for 3D images analysis and recognition based on the spatial trace transform with random scan parameters. Official registration certificate for computer programs No. 2015612257 of the Rospatent.

[16] Santalo, L. A. 2004. *Integral geometry and geometric probability*. 2nd ed. New York, NY: Cambridge University Press. 428 p.

[17] Princeton Shape Benchmark of 3D models database. Available at: http://shape.cs.princeton.edu/benchmark/ (accessed December 11, 2015).

[18] Syemov, A. A. 2014. Experimental verification of the 3D trace-transform properties. *XXI century: Past results and present problems — plus. Scientific-Methodical J. Ser.: Engineering Science. Information Technology* 03(19):83–89.

[19] Fedotov, N. G., S. V. Ryndina, and A. A. Semov. 2014. Trace transform of three-dimensional objects: Recognition, analysis and database search. *Pattern Recognition Image Anal. Adv. Math. Theory Appl.* 24(4):566–574.

[20] Fedotov, N. G., and D. A. Mokshanina. 2010. Recognition of halftone textures from the standpoint of stochastic geometry and functional analysis. *Pattern Recognition Image Anal. Adv. Math. Theory Appl.* 20(4):551–556.

[21] Fedotov, N. G., and D. A. Mokshanina. 2011. Recognition of images with complex half-tone texture. *Measurement Techniques* 53(11):1226–1232.

[22] Fedotov, N., S. Romanov, and D. Goldueva. 2013. Application of triple features theory to the analysis of half-tone images and colored textures. Feature construction along stochastic geometry and functional analysis. *Computer Inform. Sci.* 6(4):17–24.

[23] Fedotov, N. G., A. A. Syemov, and A. V. Moiseev. 2014. Intelligent capabilities hypertrace transform: Constructing features with predetermined properties. *Machine Learning Data Anal.* 1(9):1200–1214.

Received June 15, 2015

Динамическая модель организации грузоперевозок*

Л. А. Бекларян, Н. К. Хачатрян
beklar@cemi.rssi.ru; nerses@cemi.rssi.ru
ЦЭМИ РАН, Москва, Россия

Исследуется модель, описывающая процесс грузоперевозок, реализуемый в рамках ряда технологий. Рассматриваются четыре варианта модели. Первый вариант описывает транснациональные транспортные перевозки, т. е. перевозки без выделенных начальной станции отправления и конечной станции распределения грузов. Второй вариант описывает транспортные перевозки с выделенной начальной станцией отправления грузов. Третий вариант описывает транспортные перевозки с выделенными начальной станцией отправления и конечной станцией распределения грузов. Четвертый вариант описывает транспортные перевозки по круговой цепочке станций. Для всех вариантов модели изучаются режимы грузоперевозок, удовлетворяющие заданной системе контроля. Такие режимы описываются решениями типа бегущей волны для нелинейного конечно-разностного аналога уравнения параболического типа. Описаны возможные режимы грузоперевозок, исследован вопрос устойчивости стационарных режимов.

Ключевые слова: *нелинейный конечно-разностный аналог параболического уравнения; решения типа бегущей волны; устойчивость; динамические модели грузоперевозок*

DOI: 10.21469/22233792.1.13.04

Dynamic model of organization of cargo transportation*

L. A. Beklaryan and N. K. Khachatryan
CEMI RAS, 47 Nachimovky prospect, Moscow, Russia

The model describing the process of cargo transportation realized through a number of technologies is investigated. Four versions of the model are considered. The first version of the model describes the transnational cargo transportation without dedicated initial departure station and the final station cargo distribution. This version of the model describes the cargo, for which both the first and the last stations are not nodes. For such cargo transportation, it is important to describe the rule of interaction of intermediate stations. The second version of the model describes the transport cargo with a dedicated initial departure station. This version of the model describes the cargo on the long section of the route where the initial departure station is nodular. The role of the station is the most significant problem in the organization of cargo and, therefore, it has extra capacity. For such cargo transportation, it is important to describe the rule of interaction between the first station and intermediate stations, as well as the rule of interaction between intermediate stations. The third version of the model describes the cargo transportation between dedicated initial departure station and final station. This version of the model describes the cargo on the long section of the route between the two node stations. In the problem of transport cargo organization, node stations play the most important role; therefore, they have additional capacity. For such cargo transportation, it is important to describe the rules of interaction of nodal stations with intermediate stations and the rules of interaction between intermediate stations. The fourth version of the model describes the cargo

*Работа выполнена при финансовой поддержке РФФИ, проект № 15-51-05011.

transportation in a circular chain of stations. For all versions of the model, the modes of freight satisfying given control system are studied. Such regimes are described by traveling wave type solutions of nonlinear finite-difference analogue of a parabolic equation. The possible modes of freight are described, the issue of stability of stationary regimes is investigated.

Keywords: *nonlinear finite-difference analogue of a parabolic equation; cargo transportation models; traveling wave type solutions*

DOI: 10.21469/22233792.1.13.04

1 Введение

Среди проблем, связанных с работой транспорта, центральное место занимают задачи планирования и организации грузоперевозок. Впервые методы нахождения оптимального плана перевозок в нашей стране были предложены в 1930-х гг. В 1939 г. Л. В. Канторовичем [1] математически описана транспортная задача линейного программирования. Им же определен целый класс задач, близких к транспортной, предложен алгоритм для решения транспортной задачи, названный методом разрешающих множителей. В 1949 г. Л. В. Канторович и М. К. Гавурин опубликовали работу [2], в которой решалась транспортная задача с ограничениями на пропускные способности. Используя идеи общего метода Л. В. Канторовича, для решения задач линейного программирования был разработан метод потенциалов. Через год этот же метод был предложен Дж. Данцигом и Ф. Вольфом [3]. В то же время в Советском Союзе А. Л. Лурье [4] был предложен метод решения транспортной задачи путем приближения условно-оптимальными планами. В 1985 г. О. И. Авен, С. Е. Ловецкий и Г. Е. Моисеенко опубликовали работу [5], посвященную проблемам оптимального планирования и управления транспортными потоками на транспортных сетях. Были рассмотрены математические модели транспортных сетей и транспортных потоков (однородный транспортный поток, поток с усилениями и ослаблениями, поток нескольких видов на транспортной сети с ограниченной пропускной способностью звеньев, динамический транспортный поток).

Другой важной задачей, связанной с работой транспорта, является организация грузоперевозок. Такая задача рассмотрена в [6–14]. Сеть грузоперевозок на железнодорожном транспорте представляет собой большую сложную систему, моделирование которой связано с дополнительными трудностями из-за сложности сети дорог и многообразия движения поездов. При исследовании характеристик системы железнодорожных грузоперевозок в целом целесообразно использовать грубые модели, в которые вводятся существенные аппроксимации, а ряд деталей не учитывается. В то же время при детальном исследовании изолированных участков сети используется точная модель, в которой связи данного участка с другими более или менее опускаются и детально исследуется только этот участок. При этом не следует упускать из виду отклонение модели от реальной сети в первом случае и недоучет связей участков во втором. Создавать модель, которая точно представляет все детали, бессмысленно, поскольку это приводит к необоснованному усложнению процесса ее проектирования, поэтому при моделировании всегда используется ряд аппроксимаций реальных характеристик движения поездов.

Данная работа посвящена детальному изучению процесса организации грузоперевозок в целом. В ней построена и исследована динамическая модель организации грузоперевозок на протяженном участке пути с большим количеством промежуточных станций, через которые проходит грузопоток. Предполагается, что между двумя соседними станциями существует межстанционный перегонный путь, где временно может храниться часть грузов.

Емкость перегонных путей считаем неограниченной. Движение грузов происходит в одном направлении. На произвольную промежуточную станцию груз может поступать как с предыдущей станции, так и с перегонного пути, расположенного между ними. Аналогично с произвольной промежуточной станции груз может быть отправлен либо на следующую станцию, либо на перегонный путь, расположенный между ними. Рассматриваются четыре варианта модели.

2 Модель транснациональных грузоперевозок

Такая модель описывает движение грузопотока без выделенных начальной станции отправления и конечной станции распределения грузов, вследствие чего считаем, что число промежуточных станций бесконечно как в правую, так и в левую стороны. Работа всех станций состоит из приема, обработки и отправки грузов, а сами станции имеют заданную пропускную способность. Под пропускной способностью понимаем максимальный объем грузов, который может пройти через промежуточную станцию за единичный отрезок времени. Обработка грузов происходит в узлах станций. В каждый момент времени число задействованных узлов на n-й станции обозначим через $z_n(t)$. В каждом узле в течение единицы времени обрабатывается единичный объем грузов. Очевидно, что количество задействованных узлов обработки грузов при бесперебойной работе всей цепи перевозок ограничено. Максимальное количество таких узлов, обозначаемое через Δ, определяет пропускную способность станций. Организация подобных грузопотоков зависит от технологий по приему, обработке и отправлению грузов. Опишем эти технологии.

Первая технология основана на установленных нормативных правилах взаимодействия соседних станций. Для каждой станции с номером i существуют правила взаимодействия с предыдущей $(i-1)$-й станцией и последующей $(i+1)$-й станцией. Согласно правилу взаимодействия с предыдущей станцией станция с номером i увеличивает количество задействованных узлов с интенсивностью $\alpha(z_{i-1} - z_i)$, если количество задействованных узлов на ней меньше, чем на предыдущей станции. При этом грузопоток принимается с предыдущей станции. В противном случае станция с номером i уменьшает количество задействованных узлов с такой же интенсивностью и грузопоток отправляется на перегонный путь.

Согласно правилу взаимодействия с последующей станцией станция с номером i уменьшает количество задействованных узлов с интенсивностью $\alpha(z_i - z_{i+1})$, если количество задействованных узлов на ней больше, чем на следующей станции. При этом грузопоток отправляется на следующую станцию. В противном случае станция с номером i увеличивает количество задействованных узлов с такой же интенсивностью и грузопоток принимается с перегонного пути.

Первая технология не учитывает условие ограниченности пропускной способности станций. Кроме того, она не позволяет использовать весь потенциал станций. В связи с этим, наряду с первой технологией, используется и иная технология.

Вторая технология позволяет как увеличить число задействованных узлов (если оно меньше Δ), так и уменьшать (если оно превышает Δ). При этом груз принимается с перегонного пути либо отправляется на перегонный путь. Функция $\varphi(\cdot)$, задающая скорость изменения числа задействованных узлов в рамках второй технологии, имеет вид, изображенный на рис. 1.

Таким образом, с учетом работы первой и второй технологий скорость изменения числа задействованных узлов для i-й станции будет описываться дифференциальным уравне-

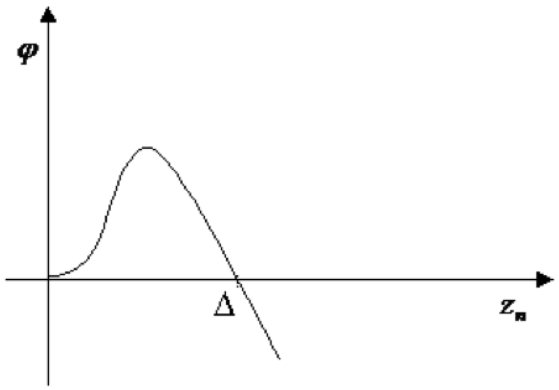


Рис. 1 Скорость изменения числа задействованных узлов (вторая технология)

нием:

$$\dot{z}_i(t) = \alpha(z_{i-1} - z_i) - \alpha(z_i - z_{i+1}) + \varphi(z_i), \quad i \in \mathbb{Z}, \quad t \in [0, +\infty). \tag{1}$$

Для грузоперевозок необходимо иметь действенную и простую систему контроля. Она заключается в том, что объемы обрабатываемых грузов для любого планового интервала времени на всех станциях должны совпадать с определенным лагом времени, единым для всех станций. Такое условие можно описать в следующем виде: существует число $\tau > 0$, не зависящее от t и i, такое, что при всех $i \in \mathbb{Z}$ и $t \in [0, +\infty)$ выполняется равенство:

$$z_i(t) = z_{i+1}(t + \tau). \tag{2}$$

Решения системы дифференциальных уравнений (1), удовлетворяющие условию (2), называются решениями типа бегущей волны. Константу τ, которая является сдвигом между моментами замеров и сравнения объемов грузов, будем называть характеристикой системы контроля. Таким образом, данная модель, описывающая процесс грузоперевозок и их систему контроля, задается счетной системой дифференциальных уравнений и условием, задающим бегущую волну:

$$\dot{z}_i(t) = \alpha z_{i-1} - 2\alpha z_i + \alpha z_{i+1} + \varphi(z_i), \quad i \in \mathbb{Z}, \quad t \in [0, +\infty); \tag{3}$$

$$z_i(t) = z_{i+1}(t + \tau), \quad i \in \mathbb{Z}, \quad t \in [0, +\infty). \tag{4}$$

Определение 1 [10]. Семейство абсолютно непрерывных функций $\{z_i(\cdot)\}_{i\in\mathbb{Z}}$, определенных на $[0, +\infty)$, называется решением системы дифференциальных уравнений (3), если при почти всех $t \in [0, +\infty)$ функции $z_i(\cdot)$ удовлетворяют этой системе. ■

Для любого $\mu \in (0, 1)$ определим банаховы пространства (пространства функций с весами)

$$\mathcal{L}^1_\mu C^{(k)}(\mathbb{R}) = \left\{ x(\cdot) : x(\cdot) \in C^{(k)}(\mathbb{R}, \mathbb{R}), \ \max_{0\leqslant r\leqslant k} \sup_{t\in R} \left\| x^{(r)}(t) e^{-\delta|t|} \right\|_{\mathbb{R}} < +\infty \right\}, \ k = 0, 1, \ldots, \quad \mu = e^{-\delta}$$

и нормой

$$\|x\|^{(k)}_\mu = \max_{0\leqslant r\leqslant k} \sup_{t\in R} \left\| x^{(r)}(t) e^{-\delta|t|} \right\|_{\mathbb{R}},$$

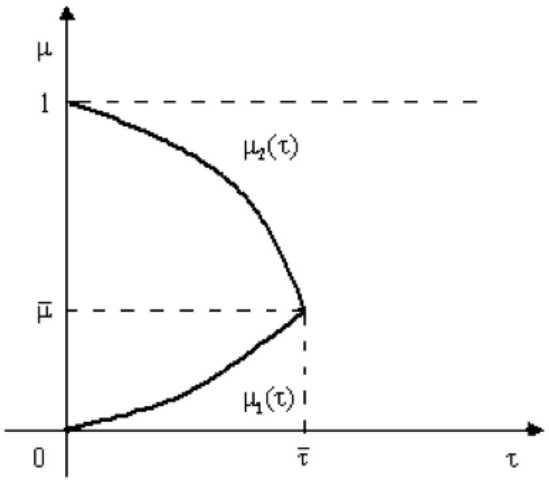


Рис. 2 Функции $\mu_1(\tau)$ и $\mu_2(\tau)$

а также векторное пространство $K^1 = \prod_{-\infty}^{+\infty} \mathbb{R}_i$, $\mathbb{R}_i = \mathbb{R}$, $i \in \mathbb{Z}$ с элементами $\varkappa = \{x_i\}_{-\infty}^{+\infty}$, $x_i \in \mathbb{R}$, $i \in \mathbb{Z}$, и со стандартной топологией полного прямого произведения.

В пространстве K^1 определим семейство гильбертовых подпространств

$$K^1_{2\mu} = \left\{ \varkappa : \varkappa \in K^1; \quad \sum_{i=-\infty}^{+\infty} |x_i|_R^2 \mu^{2|i|} < +\infty \right\}, \quad \mu \in (0,1),$$

с нормой

$$\|\varkappa\|_{2\mu} = \left[\sum_{i=-\infty}^{+\infty} |x_i|_R^2 \mu^{2|i|} \right]^{1/2}.$$

Обозначим

$$M(\tau) = \tau \max[2\alpha, L_0]$$

и рассмотрим неравенство относительно двух переменных $\tau \in (0, +\infty)$ и $\mu \in (0,1)$

$$M(\tau)[1 + 2\mu^{-1}] < \ln \mu^{-1}, \qquad \mu \in (0,1). \tag{5}$$

Множество решений неравенства (5) описывается функциями $\mu_1(\tau)$ и $\mu_2(\tau)$, изображенными на рис. 2.

Теорема 1 [10]. *Для любых начальных данных $a > 0$, $\bar{i} \in \mathbb{Z}$, начального момента времени $\bar{t} \in [0, +\infty)$ и характеристик τ, удовлетворяющих условию $0 < \tau < \bar{\tau}$, существует решение $\{z_i(\cdot)\}_{i\in\mathbb{Z}}$ уравнения (3) типа бегущей волны (условие (4)) с характеристикой τ, удовлетворяющее начальному условию $z_{\bar{i}}(\bar{t}) = a$. Более того, в таком решении для всякого $i \in \mathbb{Z}$ функция $z_i(\cdot)$ принадлежит пространству $\mathcal{L}^1_{\sqrt[i]{\mu}} C^{(0)}([0, +\infty))$ при любом $\mu \in (\mu_1(\tau), \mu_2(\tau))$. Такое решение является единственным и непрерывно зависит от начального условия a, каждая координата $z_i(\cdot)$, $i \in \mathbb{Z}$, непрерывно зависит от начального условия a как элемент пространства $\mathcal{L}^1_{\sqrt[i]{\mu}} C^{(0)}([0, +\infty))$.* ■

Система (3)–(4) имеет два стационарных решения типа бегущей волны: $\bar{z}_1 \equiv \{\cdots, 0, 0, 0, \cdots\}$, $\bar{z}_2 \equiv \{\cdots, \Delta, \Delta, \Delta, \cdots\}$. Очевидно, что такие решения принадлежат пространству $K^1_{2\mu}$ при любом $\mu \in (0,1)$.

Рассмотрим уравнение

$$\alpha\mu^2 - (2\alpha + \delta)\mu + \alpha = 0, \tag{6}$$

где $\delta = -\varphi'(\Delta)$. Решениями уравнения (6) являются $\tilde{\lambda}$, $\tilde{\tilde{\lambda}}$, причем $0 < \tilde{\lambda} < 1$, $\tilde{\tilde{\lambda}} > 1$.

Определение 2 [10]. Стационарное решение $\bar{z} = \{\bar{z}_i\}_{i\in\mathbb{Z}}$ системы уравнений (3) в фазовом пространстве $K^1_{2\mu}$, $\mu \in (0,1)$, называется устойчивым по Ляпунову, если существуют $\gamma > 0$ и $\bar{t} \geqslant 0$ такие, что для произвольного $d \in K^1_{2\mu}$, удовлетворяющего условию $\|d - \bar{z}\|_{2\mu} < \gamma$, решение $z(t)$ уравнения (3) с начальным условием $z(\bar{t}) = d$ существует; для всякого $\varepsilon > 0$ существует $0 < \sigma_1 < \gamma$ такое, что при $\|d - \bar{z}\|_{2\mu} < \sigma_1$ решение $z(t)$ уравнения (3) с начальным условием $z(\bar{t}) = d$ удовлетворяет условию $\|z(t) - \bar{z}\|_{2\mu} < \varepsilon$ для всех $t > \bar{t}$.

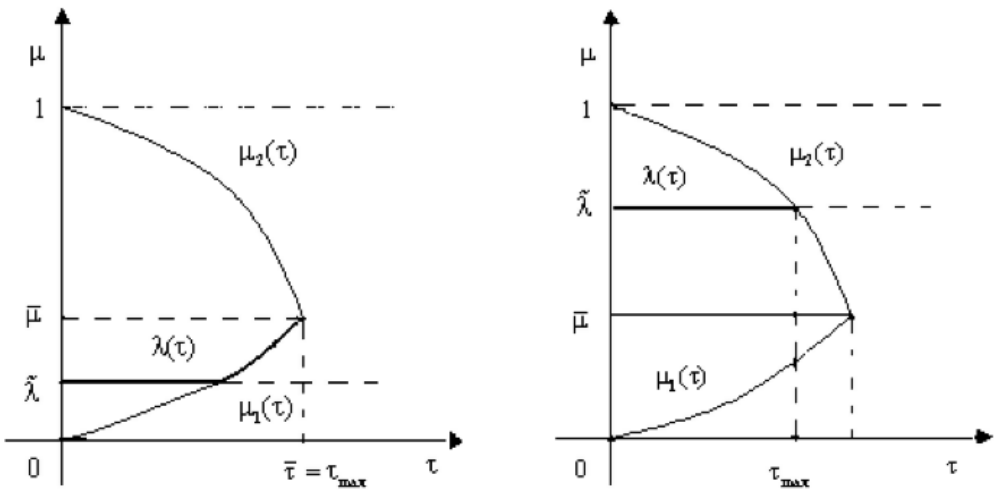
Устойчивое по Ляпунову стационарное решение $\bar{z} = \{\bar{z}_i\}_{i\in\mathbb{Z}}$ системы уравнений (3) в фазовом пространстве $K^1_{2\mu}$, $\mu \in (0,1)$, называется асимптотически устойчивым, если $\lim\limits_{t\to+\infty} \|z(t) - \bar{z}\|_{2\mu} = 0$. ■

Теорема 2 [10]. *Для любых $\alpha, \delta > 0$ и характеристик $\tau \in (0, +\infty)$ стационарное решение $\bar{z}_2 = \{\cdots, \Delta, \Delta, \Delta, \cdots\}$ уравнения (3) в фазовом пространстве $K^1_{2\mu}$, $\mu \in (\tilde{\lambda}, 1)$, является асимптотически устойчивым, а стационарное решение $\bar{z}_1 = \{\cdots, 0, 0, 0, \cdots\}$ в фазовом пространстве $K^1_{2\mu}$, $\mu \in (0,1)$, является неустойчивым.* ■

Обозначим

$$\tau_{\max} = \sup\{\tau : \tau \leqslant \bar{\tau},\ \mu_2(\tau) \geqslant \tilde{\lambda}\}.$$

На интервале $(0, \tau_{\max}]$ определяется функция $\lambda(\tau) = \max(\tilde{\lambda}, \mu_1(\tau))$, графически изображенная на рис. 3 (при $\tilde{\lambda} < \bar{\mu}$ — на рис. 3, *а* и при $\tilde{\lambda} > \bar{\mu}$ — на рис. 3, *б*).



(*а*) (*б*)

Рис. 3 Функция $\lambda(\tau)$: (а) $\tilde{\lambda} < \overline{\mu}$; (б) $\tilde{\lambda} > \overline{\mu}$

Определение 3 [10]. Стационарное решение $\bar{z} = \{\bar{z}_i\}_{i\in\mathbb{Z}}$, $\bar{z}_i = \bar{z}_{i+1}$, $i \in \mathbb{Z}$, типа бегущей волны системы уравнений (3) в фазовом пространстве $K^1_{2\mu}$, $\mu \in (0,1)$, называется устойчивым по Ляпунову среди решений типа бегущей волны с характеристикой τ, если:

оно устойчиво по Ляпунову; существуют $\gamma > 0$ и $\bar{t} \geqslant 0$ такие, что для произвольного числа d_0, удовлетворяющего условию $|d_0 - \bar{z}_0| < \gamma$, решение $z(t) = \{z_n(t)\}_{n\in\mathbb{Z}}$ системы (3)–(4) с начальным условием $z_0(\bar{t}) = d_0$ существует; для всякого $\varepsilon > 0$ существует $0 < \sigma_2 < \gamma$ такое, что из условия $|d_0 - \bar{z}_0| < \sigma_2$ следует, что решение $z(t)$ системы (3)–(4) с начальным условием $z_0(\bar{t}) = d_0$ удовлетворяет условию $||z(t) - \bar{z}||_{2\mu} < \varepsilon$ для всех $t > \bar{t}$. ■

Имеет место следующая теорема.

Теорема 3 [10]. *Для любых $\alpha, \delta > 0$ и характеристик $\tau \in (0, \tau_{\max})$ стационарное решение $\bar{z}_2 = \{\cdots, \Delta, \Delta, \Delta, \cdots\}$ системы уравнений* (3) *в фазовом пространстве $K^1_{2\mu}$, $\mu \in (\lambda(\tau), \mu_2(\tau))$ является асимптотически устойчивым среди решений типа бегущей волны с характеристикой τ.* ■

3 Модель грузоперевозок с выделенной начальной станцией отправления грузов

В предыдущем разделе была рассмотрена модель транснациональных транспортных перевозок, где предполагалось, что множество промежуточных станций бесконечно как в правую, так и в левую стороны. В данном разделе рассмотрим модель транспортных перевозок с выделенной начальной станцией отправления грузов. Итак, рассмотрим модель транспортных перевозок с начальной станцией отправления грузов $i = 0$ и большим количеством промежуточных станций $i = 1, 2, \ldots$ Так же, как и в первой модели, организация грузопотока осуществляется посредством двух технологий.

Первая технология. На станциях с номерами $i = 1, 2, ...$ действует первая технология, описанная в предыдущем параграфе. На начальной станции $i = 0$ первая технология определяется с помощью правила взаимодействия с последующей станцией и правила подачи грузов на нее, определяемая функцией $\psi(t)$, зависящей от переменной времени $t \geqslant 0$. Предполагаем, что функция $\psi(\cdot)$ является кусочно бесконечно дифференцируемой. Так как начальная станция является узловой, то естественно предположить, что она обладает большими мощностями и при необходимости на ней можно резко изменять число задействованных узлов, чего нельзя сделать на промежуточных станциях.

Вторая технология. Для произвольной станции с номером $i = 1, 2, ...$ вторая технология в точности повторяет вторую технологию, описанную в предыдущем разделе. Для начальной станции $i = 0$ вторая технология из предыдущего параграфа используется только для разгрузки, поэтому скорость изменения числа задействованных узлов обработки на начальной станции в рамках второй технологии описывается функцией $\varphi_0(t)$, зависящей от количества задействованных узлов на начальной станции, и удовлетворяет следующим условиям: на полупрямой $(-\infty, \Delta]$ тождественно равна нулю, а на полупрямой $[\Delta, +\infty)$ является убывающей функцией. Предполагаем, что функции $\varphi_0(\cdot)$ и $\varphi(\cdot)$ (определенная в предыдущем параграфе) являются бесконечно дифференцируемыми. Очевидно, что при объеме грузов на 0-й станции, не превышающем Δ, используется только первая технология.

Таким образом, с учетом работы первой и второй технологий, а также системы контроля, процесс грузоперевозок будет описываться следующей системой дифференциальных уравнений

$$\left.\begin{aligned} &\dot{z}_0(t) = \psi(t) - \alpha z_0 + \alpha z_1 + \varphi_0(z_0), \quad t \in [0, +\infty); \\ &\dot{z}_i(t) = \alpha z_{i-1} - 2\alpha z_i + \alpha z_{i+1} + \varphi(z_i), \quad i = 1, 2, \ldots, \quad t \in [0, +\infty); \\ &z_i(t) = z_{i+1}(t + \tau), \quad i = 0, 1, 2, ..., \quad t \in [0, +\infty). \end{aligned}\right\} \tag{7}$$

Класс решений системы (7) чрезвычайно узок, поэтому для описания реализуемых режимов грузоперевозок используется более широкий класс решений, которые называются квазирешениями типа бегущей волны. Эти решения являются кусочно абсолютно непрерывными, а разрывы расположены в точках, кратных характеристике системы контроля (параметр τ). Приведем точное определение.

Определение 4 [10]. Семейство кусочно абсолютно непрерывных функций $\{z_i(.)\}_0^{+\infty}$, определенных на $[0,+\infty)$, называется квазирешением типа бегущей волны с характеристикой $\tau > 0$ для системы (7), если при почти всех $t \in [0,+\infty)$ функции $z_i(\cdot)$ удовлетворяют этой системе, а разрывы расположены в точках, кратных числу τ. ■

Теорема 4 [10]. *Для любых начальных данных $a > 0$, $\bar{i} \in \{0,1,\ldots\}$, начального момента времени $\bar{t} \in [0,+\infty)$, характеристик τ, удовлетворяющих условию $0 < \tau < \bar{\tau}$ (см. рис. 2) и функций $\psi(\cdot) \in C^\infty([0,\tau],\mathbb{R})$ на полупрямой $(\tau,+\infty)$ существует единственное кусочно непрерывное продолжение функции $\psi(\cdot)$ и соответсвующее ему квазирешение $\{z_i(\cdot)\}_0^{+\infty}$ типа бегущей волны с характеристикой τ системы (7) в фазовом пространстве $K_{2\mu}^1$, $\mu \in (\mu_1(\tau),\mu_2(\tau))$, удовлетворяющее начальному условию $z_{\bar{i}}(\bar{t}) = a$. Такое квазирешение является единственным и непрерывно зависит от начального условия a и функции $\psi(\cdot)$.* ■

В содержательном плане это означает, что на всех станциях в моменты времени, кратные характеристике системы контроля, необходимо резко менять число задействованных узлов. Данная процедура требует подключения дополнительных мощностей, которые имеются только на узловой (начальной) станции. Оказывается, что достаточно лишь на начальной станции в начальный период времени резко изменить число задействованных узлов (слегка изменить функцию $\psi(\cdot)$ в норме $L_1([0,\tau],\mathbb{R})$), чтобы организовать контролируемый грузопоток с помощью определенных выше технологий (получить так называемое ε квазирешение, т. е. такое квазирешение, у которого указанные разрывы меньше ε).

Определение 5 [10]. Квазирешение типа бегущей волны с характеристикой τ называется ε-квазирешением типа бегущей волны с характеристикой τ или (ε,τ)-квазирешением, если выполняются неравенства:

$$|z_0(k\tau - 0) - z_0(k\tau + 0)| < \varepsilon, \quad k = 1,2,\ldots \ ■$$

Теорема 5 [10]. *Для любых начальных данных $a > 0$, $\bar{i} \in \{0,1,\ldots\}$, начальных моментов времени $\bar{t} \in [0,+\infty)$, характеристик τ, удовлетворяющих условию $0 < \tau < \bar{\tau}$, произвольной функции $\psi(\cdot) \in C^\infty([0,\tau],\mathbb{R})$ и произвольного $\varepsilon > 0$ существует функция $\psi_\varepsilon(\cdot) \in C^\infty([0,\tau],\mathbb{R})$, отличная от $\psi(\cdot)$ в малой окрестности точки 0 такая, что ее продолжение на $(\tau,+\infty)$ и соответствующее ему квазирешение $\{z_{i\varepsilon}(\cdot)\}_0^{+\infty}$ типа бегущей волны с характеристикой τ системы (7), удовлетворяет начальному условию $z_{\bar{i}\varepsilon}(\bar{t}) = a$, принадлежит фазовому пространству $K_{2\mu}^1$ при любом $\mu \in (\mu_1(\tau),\mu_2(\tau))$ и является (ε,τ)-квазирешением.* ■

4 Модель грузоперевозок с выделенными начальной станцией отправления и конечной станцией распределения грузов

Рассмотрим модель транспортных перевозок с начальной станцией отправления грузов $i = 0$, конечным числом промежуточных станций $i = 1,2,\ldots,m$ и конечной станцией распределения грузов $i = m+1$. Так же, как и в предыдущих моделях, организация грузопотока осуществляется посредством двух технологий.

Первая технология. На станциях с номерами $i = 0,1,2,\ldots,m$ действует технология, описанная раннее. Технология подачи грузов на начальную станцию описывается функ-

цией $\psi_1(t)$, $t \geqslant 0$. На конечной станции первая технологии определяется с помощью правила взаимодействия с предыдущей станцией и правилом распределения грузов с нее, описываемая функцией $\psi_2(t)$, $t \geqslant 0$. Предполагаем, что функция $\psi_1(\cdot)$ является кусочно бесконечно дифференцируемой, а функция $\psi_2(\cdot)$ — кусочно непрерывной.

Вторая технология. Для начальной и промежуточных станций вторая технология в точности повторяет вторую технологию, описанную в предыдущих параграфах. Вторая технология для конечной станции такая же, как для промежуточных станций.

Таким образом, с учетом работы первой и второй технологий, а также системы контроля прием и отправка грузов будут описываться следующей системой дифференциальных уравнений:

$$\left.\begin{array}{rcll} \dot{z}_0(t) &=& \psi_1(t) - \alpha z_0 + \alpha z_1 + \varphi_0(z_0)\,, & t \in [0,+\infty); \\ \dot{z}_i(t) &=& \alpha z_{i-1} - 2\alpha z_i + \alpha z_{i+1} + \varphi(z_i)\,, \quad i = 1,2,\ldots,m\,, & t \in [0,+\infty); \\ \dot{z}_{m+1}(t) &=& \alpha z_m - \alpha z_{m+1} - \psi_2(t) + \varphi(z_{m+1})\,, & t \in [0,+\infty); \\ z_i(t) &=& z_{i+1}(t+\tau)\,, \quad i = 0,1,2,\ldots,m\,, & t \in [0,+\infty). \end{array}\right\} \tag{8}$$

Класс решений системы (8) также чрезвычайно узок и для описания реализуемых режимов грузоперевозок используются квазирешения (имеются разрывы в точках, кратных характеристике системы контроля) типа бегущей волны.

Теорема 6 [10]. *Для любых начальных данных $a > 0$, $\bar{i} \in \{0,1,\ldots,m+1\}$, начального момента времени $\bar{t} \in [0,+\infty)$, характеристик τ, удовлетворяющих условию $0 < \tau < \bar{\tau}$, (см. рис. 2) и функций $\psi_1(\cdot) \in C^\infty([0,\tau],\mathbb{R})$ и $\psi_2(\cdot) \in C([0,\tau],\mathbb{R})$ существуют единственные кусочно непрерывные продолжения функций $\psi_1(\cdot)$ и $\psi_2(\cdot)$ и соответствующее им квазирешение $\{z_i(\cdot)\}_0^{m+1}$ типа бегущей волны с характеристикой τ системы (8) в фазовом пространстве $K^1_{2\mu}$, $\mu \in (\mu_1(\tau),\mu_2(\tau))$, удовлетворяющее начальному условию $z_{\bar{i}}(\bar{t}) = a$. Такое квазирешение является единственным и непрерывно зависит от начального условия a и функций $\psi_1(\cdot)$ и $\psi_2(\cdot)$.* ■

Оказывается, что так же, как и для предыдущей модели (с выделенной начальной станцией отправления грузов), с помощью резкого изменения числа задействованных узлов на начальной станции в начальный период времени можно организовать контролируемый грузопоток (получить ε-квазирешение).

Теорема 7 [10]. *Для любых начальных данных $a > 0$, $\bar{i} \in \{0,1,\ldots,m+1\}$, начальных моментов времени $\bar{t} \in [0,+\infty)$, характеристик τ, удовлетворяющих условию $0 < \tau < \bar{\tau}$, произвольных функций $\psi_1(\cdot) \in C^\infty([0,\tau],R)$ и $\psi_2(\cdot) \in C([0,\tau],R)$ и произвольного $\varepsilon > 0$ существует функция $\psi_{1\varepsilon}(\cdot) \in C^\infty([0,\tau],R)$, отличная от $\psi_1(\cdot)$ в малой окрестности точки 0, такая, что продолжения функций $\psi_{1\varepsilon}(\cdot)$ и $\psi_2(\cdot)$ на $(\tau,+\infty)$ и соответсвующее им квазирешение $\{z_{i\varepsilon}(\cdot)\}_0^{m+1}$ типа бегущей волны с характеристикой τ системы (8) удовлетворяет начальному условию $z_{\bar{i}\varepsilon}(\bar{t}) = a$, принадлежит фазовому пространству $K^1_{2\mu}$ при любом $\mu \in (\mu_1(\tau),\mu_2(\tau))$ и является (ε,τ)-квазирешением.* ■

5 Модель грузоперевозок по круговой цепочке станций

Вернемся к первому варианту модели. Напомним, что этот вариант модели описывает транснациональные транспортные грузоперевозки без выделенных начальной станции отправления и конечной станции распределения грузов. Рассмотрим частный случай такой модели, а именно: модель транспортных грузоперевозок по круговой цепочке, состоящей из n станций. Для исследования данной модели необходимо изучить решения системы (3)–(4), удовлетворяющие следующему дополнительному условию:

$$z_i(t) = z_{i+n}(t), \quad i \in Z, \quad t \in [0,\ +\infty).$$

Таким образом, данная модель описывается следующей системой:

$$\left.\begin{array}{rcll} \dot{z}_i(t) &=& \alpha z_{i-1} - 2\alpha z_i + \alpha z_{i+1} + \varphi(z_i), & i \in Z, \quad t \in [0,+\infty); \\ z_i(t) &=& z_{i+n}(t), & i \in Z, \quad t \in [0,+\infty); \\ z_i(t) &=& z_{i+1}(t+\tau), & i \in Z\,, \quad t \in [0,+\infty). \end{array}\right\} \tag{9}$$

Справедлива следующая лемма.

Лемма 1 [14]. *Если $\{\bar{z}_i(\cdot)\}_{i\in\mathbb{Z}}$ является решением системы* (9)*, то для произвольного $i \in Z$ функция $\bar{z}_i(\cdot)$ периодическая с периодом τn.* ■

Очевидно, что разрешимость системы (9) зависит от разрешимости следующей конечномерной системы:

$$\left.\begin{array}{rcll} \dot{z}_1(t) &=& \alpha z_n - 2\alpha z_1 + \alpha z_2 + \varphi(z_1), & t \in [0,+\infty); \\ \dot{z}_i(t) &=& \alpha z_{i-1} - 2\alpha z_i + \alpha z_{i+1} + \varphi(z_i), & i = 2,\ldots,n-1, \quad t \in [0,+\infty); \\ \dot{z}_n(t) &=& \alpha z_{n-1} - 2\alpha z_n + \alpha z_1 + \varphi(z_n)\,, & t \in [0,+\infty); \end{array}\right\} \tag{10}$$

$$\left.\begin{array}{rll} z_i(t) &= z_{i+1}(t+\tau), & i = 1,\ldots,n-1\,, \quad t \in [0,+\infty)\,; \\ z_n(t) &= z_1(t+\tau)\,, & t \in [0,+\infty)\,. \end{array}\right\} \tag{11}$$

Итак, согласно лемме 1, если система (10)–(11) имеет решение, то оно будет периодическим с периодом τn. Одним из таких решений является стационарное решение $(\Delta, \Delta, \ldots, \Delta)$. Для выявления других решений (если они существуют) изучим все решения системы дифференциальных уравнений (10) (т. е. не только решения типа бегущей волны, удовлетворяющие условиям (11)).

Теорема 8 [14]. *Для произвольных $\alpha > 0$ и $\delta > 0$ всякое решение системы дифференциальных уравнений* (10) *с координатами начального значения, бóльшими 0, ограничено. Более того, каждая координата решения снизу ограничена нулем, а сверху асимптотически ограничена значением Δ.* ■

Теорема 9 [14]. *Стационарное решение $(\Delta, \Delta, \ldots, \Delta)$ системы дифференциальных уравнений* (10) *локально устойчиво по Ляпунову по первому приближению.* ■

Для определения области устойчивости указанного решения система дифференциальных уравнений (10) была решена численно с помощью метода Рунге–Кутта 4-го порядка. Результаты численных экспериментов сформулируем в виде следующего утверждения.

Утверждение 1 [14]. *Для любых $\alpha > 0$ и $\delta > 0$ областью устойчивости стационарного решения $(\Delta, \Delta, \ldots, \Delta)$ системы дифференциальных уравнений* (10) *является положительный ортант, т. е. всякое решение системы дифференциальных уравнений* (10) *с положительными координатами начального значения сходится к стационарному решению $(\Delta, \Delta, \ldots, \Delta)$.* ■

Из теоремы 8 и утверждения 1 следует, что всякое решение системы дифференциальных уравнений (10) с положительными координатами начального значения ограничено и, более того, сходится к стационарному решению $(\Delta, \Delta, \ldots, \Delta)$. Следовательно, других периодических решений, кроме стационарного решения $(\Delta, \Delta, \ldots, \Delta)$, система дифференциальных уравнений (10) с положительными координатами начального значения не имеет. Это, в свою очередь, означает, что система (10)–(11) с положительными координатами начального значения, кроме стационарного решения $(\Delta, \Delta, \ldots, \Delta)$, не имеет других решений. Таким образом, исходная система (9) с положительными координатами начального значения имеет единственное решение типа бегущей волны, а именно: стационарное решение $(\Delta, \Delta, \ldots, \Delta)$.

Литература

[1] *Канторович Л. В.* Математические методы организации и планирования производства. — Л.: Изд-во ЛГУ, 1939. 68 с.

[2] *Канторович Л. В., Гавурин М. К.* Применение математических методов в вопросах анализа грузопотоков // Проблемы повышения эффективности работы транспорта: Сб. научн. статей. — М.: Изд-во АН СССР, 1949. С. 110–138.

[3] *Данциг Дж., Вольфф Ф.* Алгоритм разложения для задач линейного программирования // Математика: Сб. переводов, 1964. Т. 8. № 1. С. 151–160.

[4] *Лурье А. Л.* Алгоритм решения сетевой транспортной задачи с ограничением пропускных способностей методом условно-оптимальных планов // Мат-лы Конф. по опыту и перспективам применения математических методов и ЭММ в планировании. — Новосибирск, 1962. С. 3–13.

[5] *Авен О. И., Ловецкий С. Е., Моисеенко Г. Е.* Оптимизация транспортных потоков. — М.: Наука, 1985. 166 с.

[6] *Козовский И. Г.* Рационализация перевозок грузов на железных дорогах. — М.: Транспорт, 1977. 280 с.

[7] *Галабурда В. Г.* Совершенствование технологии перевозок и увеличение пропускной способности железных дорог. — М.: МИИТ, 1983. 124 с.

[8] *Галабурда В. Г.* Оптимальное планирование грузопотоков. — М.: Транспорт, 1985. 256 с.

[9] *De Jong G., Gunn H. F., Walker W.* National and international freight transport models: An overview and ideas for further development // Transport Rev., 2004. Vol. 24. No. 1. P. 103–124.

[10] *Beklaryan L. A., Khachatryan N. K.* Traveling wave type solutions in dynamic transport models // Functional Differential Equations, 2006. Vol. 13. No. 2. P. 125–155.

[11] *Рубцов А. О., Тарасов А. С.* Моделирование железнодорожных перевозок на территории России // Тр. Института системного анализа Российской академии наук, 2009. Т. 46. С. 274–278.

[12] *Yamada T., Russ B. F., Castro J., Taniguchi E.* Designing multimodal freight transport networks: A heuristic approach and applications // Transportation Sci., 2009. Vol. 43. No. 2. P. 129–143.

[13] *Левин Д. Ю.* Моделирование процессов перевозки // Мир транспорта, 2010. Т. 8. № 5(33). С. 48–55.

[14] *Бекларян Л. А., Хачатрян Н. К.* Об одном классе динамических моделей грузоперевозок // Журнал вычислительной математики и математической физики, 2013. Т. 53. № 10. С. 1649–1667.

Поступила в редакцию 15.06.2015

References

[1] Kantorovich, L. V. 1939. *Mathematical methods of organizing and planning production*. Leningrad: Leningrad State University. 68 p. (In Russian.)

[2] Kantorovich, L. V., and M. K. Gavurin. 1949. Application of mathematical methods to the analysis of freight flows. *Problems of raising the efficiency of transport performance*. Moscow: USSR Acad. Sci. 110–138. (In Russian.)

[3] Dantzig, G. B., and P. Wolfe. 1964. Decomposition principle for linear programs. *Oper. Res.* 8:101–111.

[4] Lurie, A. L. 1962. Algorithm for solving the restricted-capacity network transportation problem with by the method of conditionally optimal plans. *Conference on Experience and Prospects of Applying Mathematical Methods and Computers in Planning Proceedings*. Novosibirsk. 3–13. (In Russian.)

[5] Aven, O. I., S. E. Lovetskii, and G. E. Moiseenko. 1985. *Optimization of traffic flows*. Moscow: Nauka. 166 p. (In Russian.)

[6] Kozovskii, I. G. 1977. *Improvement of railroad good transportation*. Moscow: Transport. 280 p. (In Russian.)

[7] Galaburda, V. G. 1983. *Improvement of transportation techniques and increase in railroad traffic capacity*. Moscow: Mosk. Inst. Inzh. Transporta. 124 p. (In Russian.)

[8] Galaburda, V. G. 1985. *Optimal planning of good transportation*. Moscow: Transport. 256 p. (In Russian.)

[9] De Jong, G., H. F. Gunn, and W. Walker. 2004. National and international freight transport models: An overview and ideas for further development. *Transport Rev.* 24(1):103–124.

[10] Beklaryan, L. A., and N. K. Khachatryan. 2006. Traveling wave type solutions in dynamic transport models. *Functional Differential Equations* 13(2):125–155.

[11] Rubtsov, A. O, and A. S. Tarasov. 2009. Modeling of rail transport in Russia. *Proceedings of ISA RAS* 46:274–278. (In Russian.)

[12] Yamada, T., B. F. Russ, J. Castro, and E. Taniguchi. 2009. Designing multimodal freight transport networks: A heuristic approach and applications. *Transportation Sci.* 43(2):129–143.

[13] Levin, D. Yu. 2010. Modeling of railway freightage. *World of Transport and Transportation* 8(5(33)):48–55. (In Russian.)

[14] Beklaryan, L. A., and N. K. Khachatryan. 2013. On one class of dynamic transportation models. *Comp. Math. Math. Phys.* 53(10):1649–1667.

Received June 15, 2015

Использование Радон и Фурье преобразований растровых изображений для описания и отслеживания заданных объектов

Е. А. Новиков, М. А. Падалко
eugen@novikov.de, padalkom@gmail.com
Институт прикладной математики и информационных технологий БФУ им. И. Канта, Калининград, Россия

Как правило, существующие на сегодняшний день алгоритмы описания и идентификации объектов нацелены на решение задач распознавания определенного типа объектов в заданных условиях. Однако поиск универсального или более обобщенного подхода к решению данной задачи остается интересной проблемой с точки зрения академических исследований и перспективным с точки зрения практической реализации. Предлагаемый подход позволяет производить идентификацию изображений объектов по широкому спектру признаков. Метод представлен в виде общего описания алгоритма и результатов экспериментальной проверки его эффективности. Основная задача разработки метода — быстрая и качественная обработка графических данных в виде динамических изображений или видеопотоков. Доступные для сравнения методы используются преимущественно для поиска объектов в статических изображениях, в то время как авторский метод в первую очередь нацелен на работу с видеопотоками. Общедоступных видеоматериалов и данных по их обработке аналогичными методами для сравнительного анализа на момент написания статьи не найдено. Рассматриваемый метод предлагает новый способ получения набора ключевых признаков образа и функцию для их сравнения. Он основывается на применении комбинации классических методов прямого преобразования Радона к матрице изображения, одномерного преобразования Фурье к полученным интегральным проекциям и статистического анализа интегральных коэффициентов Фурье, рассматриваемых в качестве основных дескрипторов объектов изображения.

Ключевые слова: *компьютерное зрение; машинное зрение; анализ изображения; поиск заданного объекта; мониторинг видео*

DOI: 10.21469/22233792.1.13.05

The use of Radon and Fourier transformations of raster images for description and tracking of predefined objects

E. A. Novikov and M. A. Padalko
Immanuel Kant Baltic Federal University, Institute of Applied Mathematics and Information Technologies, 14 A. Nevskogo st., Kaliningrad, Russia

As a rule, currently existing algorithms for describing and objects identification are aimed at solving the problem of definite types of objects in desired condition. However, search for all-in-one or more general approach to it is still quite interesting in the context of academic researches and from the perspective of implementation. Proposed approach allows the object identification on a wide range of characteristics. The method is presented in the form of a general algorithm description and results of the experimental test of its efficiency. The main task of method development is fast and quality processing of graphical data as dynamic images

or video streams. Available for comparison methods are mainly used for object search in freeze-frame images while the presented method is primarily aimed at working with video stream. There are no generally available videos or data on their processing with similar methods for comparative study as of this paper. Considered method allows one a new way of getting a key features set of the images and a function for their comparison. It is based on a combination of classical methods of direct Radon conversion of image matrix, one-dimensional Fourier conversion to the corresponding projections, and statistical analysis of integral Fourier coefficients, considered as objects features of images.

Keywords: *computer vision; machine vision; image analysis; object detection; video monitoring*

DOI: 10.21469/22233792.1.13.05

1 Введение

На сегодняшний день абсолютное большинство методов и технических средств обработки изображений предназначены для решения узкого класса прикладных задач и относительно надежно справляются с ними в рамках своего класса, однако в условиях среды с организованными помехами надежность существующих методов резко снижается. При этом временны́е затраты на обработку и принятие решения практически всегда значительно превышают требуемые в рамках поставленной задачи.

Стандартная схема обработки информации при решении задачи распознавания представлена на рис. 1.

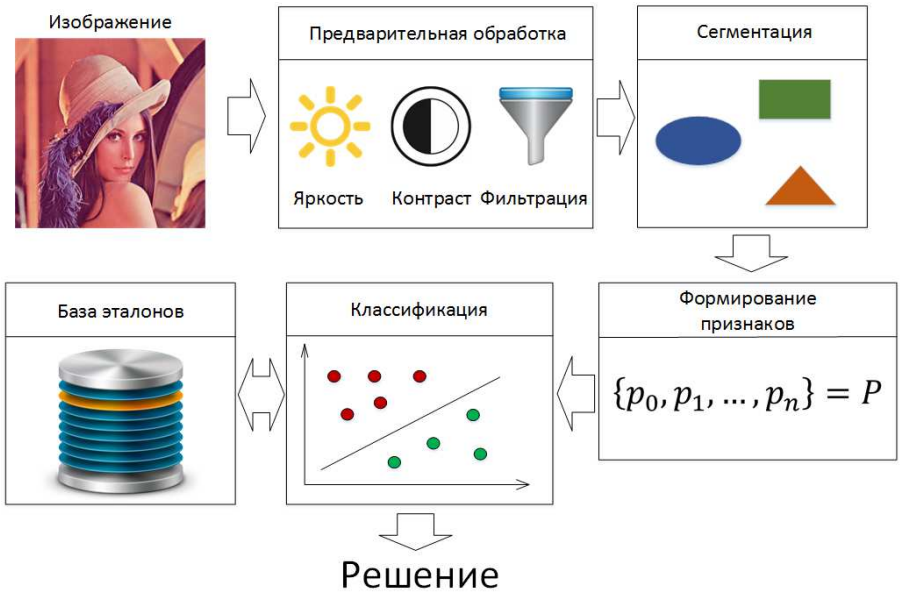


Рис. 1 Стандартная схема обработки информации при решении задачи распознавания

На данный момент для этапов сегментации и формирования признаков не существует строгого теоретического решения, так как в условиях наличия помех эти задачи являются некорректными. И поэтому качественное решение задачи распознавания изображений на указанных этапах не представляется возможным, особенно в режиме реального масштаба времени [1].

Ярким примером этой ситуации является отсутствие надежных алгоритмов для решения практических задач, таких как:

– идентификация личности по фотографии;
– анализ видеопотоков в системах мониторинга для таможни, охраны и т. п.;
– анализ рукописных текстов и (даже) текстов с гостированными шрифтами.

Рассмотрим авторский метод, реализующий новый подход к задаче распознавания образов. Новизна заключается в прямом получении ключевых параметров исследуемого изображения через последовательное применение преобразования Радона и преобразования Фурье.

2 Предлагаемый метод

2.1 Математический аппарат

В основе метода лежат известные преобразования Радона и Фурье.

Рассмотрим преобразование Радона функции двух переменных, так как именно этот случай используется в предлагаемом алгоритме распознавания.

Пусть $f(x, y)$ — функция двух действительных переменных, определенная на всей плоскости и достаточно быстро убывающая на бесконечности (так, чтобы соответствующие несобственные интегралы сходились). Тогда преобразованием Радона функции $f(x, y)$ называется функция

$$R(s, \alpha) = \int\limits_{-\infty}^{\infty} f(s\cos\,\alpha - z\sin\,\alpha, s\sin\,\alpha + z\cos\,\alpha)\,dz\,.$$

Геометрический смысл преобразования Радона — это интеграл от функции вдоль прямой, перпендикулярной вектору $\boldsymbol{n} = (\cos\,\alpha, \sin\,\alpha)$ и проходящей на расстоянии s (измеренном вдоль вектора $\boldsymbol{n}$, с соответствующим знаком) от начала координат (рис. 2) [2].

Напомним, что тригонометрическим рядом Фурье функции $f \in L_2$ с периодом T называют функциональный ряд вида

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty}\left(a_k\,\cos\frac{k2\pi x}{T} + b_k\,\sin\frac{k2\pi x}{T}\right).$$

Синусы и косинусы рядов Фурье можно рассматривать как математические резонаторы, обладающие возможностью обнаруживать присутствие определенных частот в сигнале и при наличии таковых — получать количественную оценку их индивидуального вклада в общий смешанный сигнал. Суммы считаются, как интегралы от $f(x)\cos kx\,dx$ и $f(x)\,\sin kx\,dx$ в границах периода T.

Деление интегралов на T дает нормализованные резонансные отклики, так называемые Фурье коэффициенты — два отклика a_k и b_k для каждого резонансного теста с частотой k:

$$a_k = \frac{1}{T}\int\limits_{x=T_0}^{T_0+T} f(x)\,\cos\left(kx\,\frac{2\pi}{T}\right)\,dx\,;\quad b_k = \frac{1}{T}\int\limits_{x=T_0}^{T_0+T} f(x)\,\sin\left(kx\,\frac{2\pi}{T}\right)\,dx\,.$$

Расчет значений a_k и b_k из дискретного потока данных следует правилам [3]:

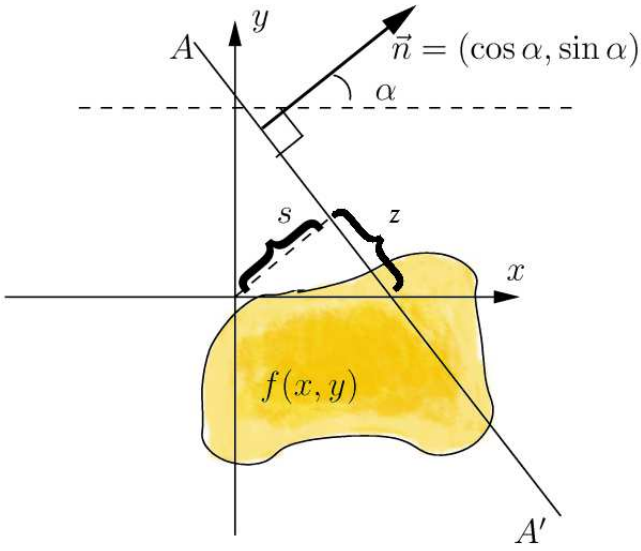


Рис. 2 Двумерное преобразование Радона. В данном случае $R(s, \alpha)$ есть интеграл от $f(x, y)$ вдоль прямой AA'

Дан сигнал $f(x)$ в форме массива длиной N; N представляет период T, и каждый резонатор записывается в виде массива длиной N. Суммы от 0 до $(N-1)$ заменяют интегралы от T_0 до (T_0+T):

$$a_k = \frac{1}{N}\sum_{x=0}^{N-1} f(x)\,\cos\left(kx\,\frac{2\pi}{N}\right); \quad b_k = \frac{1}{N}\sum_{x=0}^{N-1} f(x)\,\sin\left(kx\,\frac{2\pi}{N}\right).$$

2.2 Общий алгоритм

Для начала рассмотрим общий алгоритм, лежащий в основе данного метода. Его можно условно разбить на четыре этапа:

1) сбор входных данных. На вход подается изображение-источник, в нем задается искомая область, таким образом получается изображение-образец. Также на вход подается изображение, в котором будет осуществлен поиск данного образца — изображение-тест. Задаются значения следующих параметров:
   а) количество *интегральных проекций Радона* (ИПР);
   б) ширина спектра Фурье-резонаторов для каждой ИПР;
   в) количество первых вычисляемых *Фурье-коэффициентов* (ФК) изображений-теста и изображений-образца (можно воспринимать этот параметр как количество учитываемых при сравнении ФК, однако фактически ненужные ФК не вычисляются с целью экономии ресурсов компьютера);
   г) допустимое отклонение каждого ФК тестируемого изображения от соответствующего ФК образца;
   д) допустимое количество существенно отличающихся ФК в полном наборе данных для каждого тестируемого изображения от соответствующих ФК образца;
2) построение схемы для вычисления ИПР и массива матриц Фурье-резонаторов согласно заданным на первом этапе параметрам (размерам области образца и значений параметров а и б). Эти схема и массив матриц будут одинаковыми для всех соразмерных изображений и впоследствии будут применяться при вычислении конкретных значений для каждого обрабатываемого изображения;

3) обработка данных. С помощью построенной схемы получаются ИПР для изображения-образца, затем с использованием массива матриц резонаторов (и с учетом параметра в) вычисляются ФК для каждой ИПР. Таким образом, получаем массив массивов ФК для изображения-образца. Подобный набор данных для каждого изображения будем называть *проекционными Фурье-данными* (ПФД).
   Затем осуществляется проход по всему изображению-тесту со смещением в один пиксель (способы обхода изображения-теста могут варьироваться в зависимости от конкретной задачи). Выделяя таким образом все области соответствующего (изображению-образцу) размера, вычисляются ПФД для каждой области. Посредством специального алгоритма (использующего параметры г и д) они сравниваются с ПФД изображения-образца, и если тестируемая область считается достаточно похожей на изображение-образец, то параметры этой области дописываются в специальный массив результатов;
4) вывод результатов. В соответствующей вкладке графического интерфейса программы выводится изображение, в котором на белом фоне помещен наиболее близкий к изображению-образцу фрагмент из изображения-теста.

Таким образом, с помощью этого алгоритма можно судить о схожести двух изображений или, например, искать некоторый объект в большом изображении или видеопотоке. Можно регулировать «грубость» работы алгоритма, изменяя входные параметры и, следовательно, детальность учитываемых свойств искомого объекта.

Вычисленный результат работы описанного алгоритма практически аналогичен результату двумерного преобразования Фурье, так как последовательное применение преобразования Радона и одномерного преобразования Фурье эквивалентны [4]. Однако описанный метод требует для обработки матрицы изображения примерно в 2 раза меньше вычислений, чем двумерное преобразование Фурье [5].

2.3 Получение интегральных проекций изображения

В описании общего алгоритма говорилось, что для построения ИПР используется схема, с помощью которой будут вычисляться конкретные значения в ИПР всех соразмерных изображений. Подробно рассмотрим алгоритм построения этой схемы. Для этого необходимы лишь три параметра: ширина и высота изображения и количество выстраиваемых проекций.

Для построения всех ИПР достаточно равномерно пройти по всем (согласно введенному параметру) углам в диапазоне $[-\pi, \pi)$ [4]. С этой целью вычисляется величина шага угла по формуле: $\mathrm{dAngle} = \pi/\mathrm{projAmt}$, где projAmt — количество выстраеваемых проекций. Затем для каждого номера проекции вычисляется значение соответствующего ему угла ($\mathrm{currAngle} = \mathrm{dAngle} \cdot \mathrm{projNumb} - \pi/2$, где projNumb — номер текущей проекции) и вызывается функция, строящая схему проекции для текущего угла.

Эта функция получает на вход ширину и высоту матрицы изображения и угол направления построения проекции. В основе механизма «сбора» точек матрицы в соответствующий элемент (являющийся массивом) схемы проекции изображения лежит принцип линейной интерполяции в виде модифицированного алгоритма Брезенхема: в элемент схемы проекции добавляются координаты элементов матрицы изображения, ближе всего интерполирующие ломаной прямую, проходящую через элементы этой матрицы. Рассмотрим подробно механизм построения такой схемы.

Напомним, что все возможные значения текущего угла (currAngle) могут быть только в диапазоне $[-\pi, \pi)$. Разобьем этот диапазон на четыре сектора: $[-\pi/2, -\pi/4)$, $[-\pi/4, 0)$,

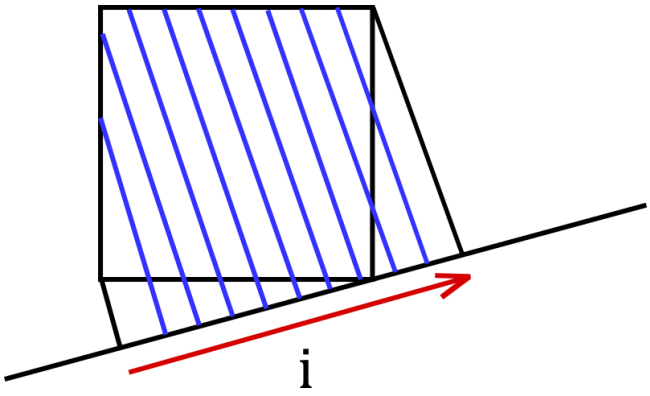


Рис. 3 Линии направлений построения проекции

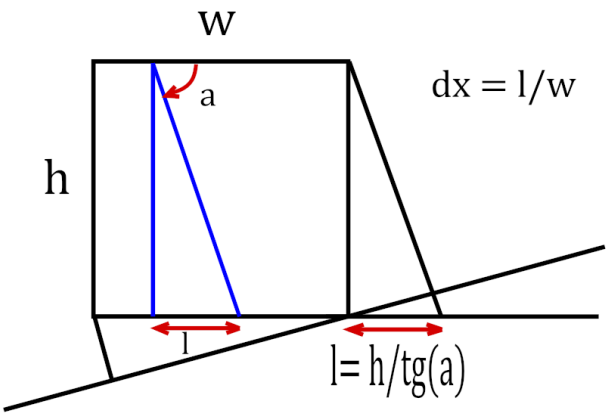


Рис. 4 Вычисление добавочной ширины и величины шага по x

$[0, \pi/4)$, $[\pi/4, \pi/2)$. Не умаляя общности, можно рассмотреть алгоритм построения схемы проекции для угла из первого сектора ($[-\pi/2, -\pi/4)$). Будем проходить по всем значениям i, для которых вдоль линии направления построения проекции будут содержаться элементы матрицы изображения (рис. 3).

Для этого сначала вычислим добавочную ширину $l = h/\operatorname{tg}(\alpha)$, где h — высота матрицы; α — угол, задающий направление построения проекции (он же = currAngle). Затем вычислим величину шага по x: $dx = l/w$, где w — ширина матрицы (рис. 4).

Для каждого значения i (с шагом 1) из достроенного отрезка $[0, w + l]$, перебирая все значения y из отрезка $[0, h]$, будем вычислять x-координату элемента матрицы, лежащего на соответствующей линии направления проекции, по формуле $x = i + \lfloor ydx \rfloor$ (где $\lfloor a \rfloor$ — «пол» числа a). Затем производится проверка выхода получившейся координаты за границы матрицы ($0 \leqslant y \leqslant h$, $0 \leqslant x \leqslant w$), и, если эта проверка пройдена успешно, получившаяся координата дописывается в i-й элемент проекции (являющийся массивом). Этот процесс отражен на рис. 5.

Общую формулу для сектора $[-\pi/2, -\pi/4)$ можно записать так:

$$P(i) = \{(x, y) : \ \forall y \in \mathbb{Z} \cap [0, h]\,, \ x = i + \lfloor y\, dx \rfloor\,, \ 0 \leqslant x \leqslant w\}\,.$$

Для других секторов принцип построения схемы проекции будет тот же, с точностью до знаков и обозначений x и y.

Таким образом, координатам матрицы изображения ставятся в соответствие номера элементов проекции. Это соответствие и позволяет построить схему проекции изображе-

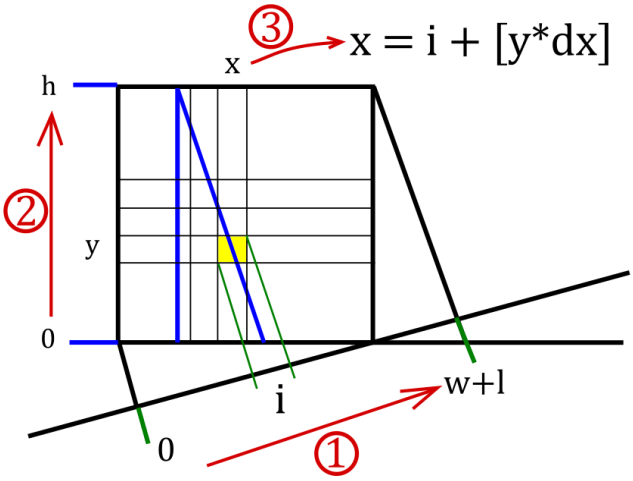


Рис. 5 Механизм отображения элемента матрицы в соответствующую ячейку проекции

ния. Отметим, что направление обхода элементов проекции всегда будет производиться слева направо.

После того как описанная функция построения схемы проекции выполнится для каждого необходимого угла, схема будет достроена и записана в память компьютера. В дальнейшем программа будет использовать эту схему для вычисления ИПР любого изображения подходящего размера: сначала один раз для изображения-образца, потом, в среднем случае, значительно большее количество раз для каждой области-теста из изображения-теста, причем вычисляться массивы ИПР будут за один «проход» по матрице обрабатываемого изображения. Заметим также, что на практике элементами каждой ИПР будут являться трехкомпонентные точки (RGB) и суммирование вдоль линии направления угла построения проекции также будет производиться покомпонентно.

Необходимость использования описанной схемы обусловлена тем фактом, что выборка из памяти по одномерному индексу выполняется быстрее, чем целочисленный расчет методом Брезенхема. Благодаря использованию схемы, производительность на этом этапе вычислений увеличилась практически вдвое для случая $80 \times 80 = 6400$ областей-тестов размером $20 \times 20 = 400$ пикселей и примерно в полтора раза для случая $260 \times 260 = 67600$ областей-тестов размером $40 \times 40 = 1600$ пикселей.

2.4 Получение Фурье-коэффициентов для интегральных проекций

Рассмотрим алгоритм построения массива матриц Фурье-резонаторов для соразмерных изображений. Для этого необходимо знать не только размеры изображения и значение параметра компрессии при Фурье-трансформации, но и длины всех ИПР обрабатываемого изображения, значение которых можно получить из схемы для вычисления ИПР. Поэтому построение массива матриц Фурье-резонаторов начинается сразу по завершении построения этой схемы.

Длина массива матриц Фурье-резонаторов будет соответствовать количеству преобразуемых последовательностей, а размеры каждой матрицы резонаторов будут зависеть от длины последовательности и от ширины спектра Фурье-резонаторов. Одно измерение матрицы — это длина последовательности ($\text{projection}_i.\text{length}$). Количество резонаторов — другое измерение матрицы — будет равно «потолку» от длины последовательности, умноженной на коэффициент ширины спектра:

$$\text{resonatorsAmt}_i = \lceil \text{projection}_i.\text{length} \cdot \text{spectrumWidth} \rceil .$$

Так же, как и в случае со схемой для построения ИПР, массив матриц Фурье-резонаторов записывается в память компьютера и будет использоваться в дальнейшем.

Как говорилось ранее, процесс получения ПФД происходит в два этапа. Сначала с использованием схемы для построения ИПР получают массив проекций, после чего с помощью массива матриц Фурье-резонаторов вычисляются массивы ФК для каждой из таких проекций, таким образом получают ПФД.

На втором этапе дополнительно используется параметр, задающий количество вычисляемых первых ФК для каждой ИПР изображения (considetationPercent). Осмысленное изменение этого параметра так же, как и в случае параметра компрессии, может способствовать решению проблемы шумоподавления в задаче распознавания образов. Влияние этого параметра, по сути, состоит в том, что при преобразовании последовательности в ФК не сохраняются последние коэффициенты, а ведь именно в последних ФК, отвечающих самым высоким частотам, как правило, содержатся шумы [6].

Таким образом, при вычислении ПФД из массива ИПР изображения достигается эффект низкочастотной Фурье-фильтрации, широко применяемый в задачах радиофизики и теории обработки сигналов [6]. Или же, если для анализа использовать самые первые ФК (т. е. значительно уменьшить значение параметра вычисляемых первых ФК), можно судить о некоторых базовых свойствах изображения.

2.5 Сравнение Фурье-коэффициентов для интегральных проекций двух изображений

Благодаря использованию описанных выше алгоритмов получаются ПФД для изображения-образца, затем, «проходя» по всему изображению-тесту, выделяются соразмерные изображению-образцу области и вычисляются ПФД для каждой такой области, после чего производится сравнение ПФД изображения-образца и области-теста.

Напомним, что представляют собой ПФД. Это массив, в котором для каждой ИПР изображения хранится некоторый массив ФК, построенный с учетом заданных параметров.

Идея метода сравнения ПФД двух изображений состоит в том, чтобы подсчитать среднее количество существенно отклоненных ФК из ПФД тестируемого изображения от ФК из ПФД изображения-образца и затем, в зависимости от этой величины, принимать решение об уровне схожести сравниваемых изображений.

Этот метод использует два последних параметра («г» и «д») из списка, упоминавшегося выше.

Первый из них — допустимый процент отклонения каждого ФК тестируемого изображения от соответствующего ФК образца, т. е. если фактическая величина отклонения ФК тестируемого изображения окажется больше значения этого параметра, то этот ФК будет считаться существенно отличающимся от соответствующего ФК образца.

Фактический процент отклонения одного ФК вычисляется как отношение значения разницы ФК тестируемого изображения от соответствующего ФК образца к значению этого ФК образца:

$$\mathrm{aDiv}_c = \frac{|\mathrm{sample}_i.a_x.c - \mathrm{test}_i.a_x.c|}{|\mathrm{sample}_i.a_x.c|}; \quad \mathrm{bDiv}_c = \frac{|\mathrm{sample}_i.b_x.c - \mathrm{test}_i.b_x.c|}{|\mathrm{sample}_i.b_x.c|}.$$

Здесь и далее i — индекс номера проекции, x — индекс номера элемента в массиве ФК, соответствующем одной проекции (ИПР после преобразования Фурье), а c принимает значения из $\{R, G, B\}$ в зависимости от цвета компоненты, для которой вычисляется отклонение.

Затем вычисляется среднее арифметическое для всех отклонений ФК каждой компоненты цвета (RGB) и типа ФК (a и b). Укажем общую формулу для подсчета фактического отклонения одного элемента ПФД:

$$\text{eleDiv}_{i\,x} = \frac{1}{6} \sum_{c=\{R,G,B\}} (\text{aDiv}_c + \text{bDiv}_c),$$

где aDiv_c и bDiv_c вычислены по формулам, указанным выше, для соответствующих значений i и x.

Далее производится подсчет количества существенно отклоненных ФК в каждой ИПР (strongDivAmt_i) в зависимости от введенного параметра допустимого процента отклонения. После этого вычисляется доля существенно отклоненных элементов массива ФК от их общего количества:

$$\text{strongDivAmtAvg}_i = \frac{\text{strongDivAmt}_i}{\text{sample}_i.\text{length}},$$

и затем считается доля существенно отклоненных ФК во всех ПФД для сравниваемого изображения:

$$\text{strongDivAmtAvgSummary} = \frac{1}{\text{projAmt}} \sum_{i=0}^{\text{projAmt}-1} \text{strongDivAmtAvg}_i\,.$$

Это значение будем называть отклонением (от образца) ПФД для изображения.

После получения этого значения нужно принять решение о том, считать ли тестируемое изображение достаточно похожим на образец. И в этот момент играет роль второй из управляющих параметров метода сравнения — допустимое количество существенно отличающихся ФК в ПФД для изображения. С ним сравнивается полученное для каждого изображения отклонение ПФД, и если оно меньше заданного параметра, то тестируемое изображение считается достаточно похожим на изображение-образец и информация о нем добавляется в специальный массив результатов поиска. В эту информацию входят координаты найденного похожего изображения и значение отклонения его ПФД.

По завершении процесса поиска схожих изображений массив результатов сортируется по неубыванию значения отклонения ПФД. За счет этого наиболее похожая на образец область оказывается на первой позиции в массиве результатов. Используя информацию о положении этой области, программа выводит соответствующую часть изображения-теста в качестве результата поиска образа из изображения-источника.

Суммарную асимптотическую вычислительную сложность алгоритма можно выразить формулой:

$$f\,(n,m,x,y,p,s,c) = \Theta\left((m-x)\,(n-y)\left(x^2+y^2\right)\left(p + 12\left(x^2+y^2\right)\frac{s\;c}{p}\right)\sum_{k=0}^{p/4} tg\left(k\,\frac{\pi}{p}\right)\right),$$

где n, m — размеры изображения-теста; x, y — размеры изображения-образца; p — количество ИПР (projAmt); s — ширина спектра Фурье-резонаторов (spectrumWidth); c — количество первых вычисляемых ФК (considerationPercent).

3 Результаты

Рассмотрим результаты, полученные с помощью программного прототипа, реализующего описанный метод распознавания образов. Как отмечалось ранее, очень большое

значение имеют параметры, задаваемые перед запуском поиска. Именно от них зависит качественная эффективность работы алгоритма. Оптимальный выбор набора параметров зависит от конкретной задачи. Ниже приведены примеры таких задач.

Для оценок ошибок будет дополнительно использоваться модификация программного прототипа, в качестве результата отображающая не только наиболее похожее на образец изображение, но и все изображения, сочтенные схожими (содержащиеся в массиве с найденными областями). На основе полученных результатов будут делаться выводы о том, какие значения управляющих параметров следует устанавливать в зависимости от конкретного изображения, выбранного в качестве образца.

3.1 Простые образы

Для начала исследуем возможности алгоритма к распознаванию простейших геометрических форм (рис. 6) и изучим особенности решения этой задачи.

Рис. 6 Простейшие геометрические формы

В качестве изображения-теста будет использоваться то же изображение с добавленными шумами (рис. 7).

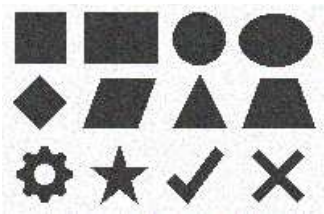


Рис. 7 Зашумленное изображение простейших геометрических форм

Начнем, например, с поиска квадрата. Образцом выбран серый квадрат (без белой границы). В подписях к рисункам с результатами поисков будут указываться использованные управляющие параметры в последовательности, изложенной в описании общего алгоритма.

Как и ожидалось, в массив результатов попал сам квадрат и прямоугольник (рис. 8). Произошло это потому, что прямоугольник содержит в себе искомый образ (квадрат), и, соответственно, он успешно распознается и попадает в массив результатов. Причем этот эффект не является ошибкой второго рода.

Теперь будем искать шестеренку. Поскольку ее форма достаточно необычна, то при правильном выборе управляющих параметров поиска найдется только она сама (рис. 9).

Однако заметим, что, как говорилось ранее, при неверном задании параметров могут возникать ошибки. Так, в том случае, если требовать слишком строгого сходства, не будет найдено вообще ни одного совпадения (на выходе будет получено пустое изображение —

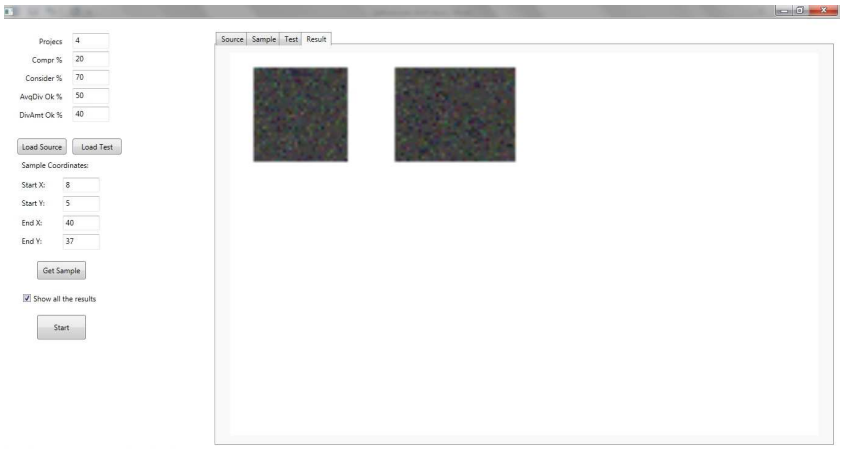


Рис. 8 Результат поиска квадрата (4, 20, 70, 50, 40)

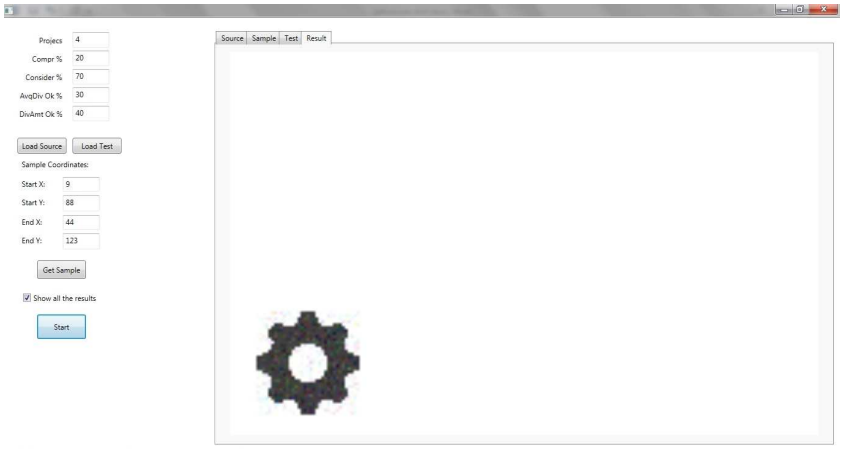


Рис. 9 Результат поиска шестеренки (4, 20, 70, 30, 40)

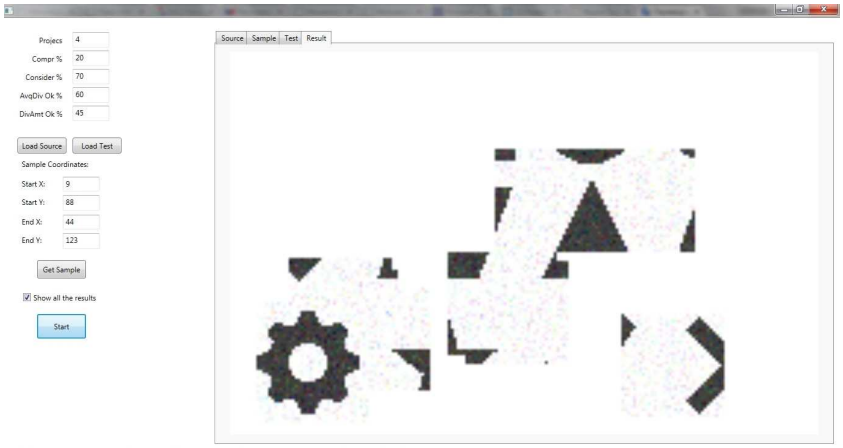


Рис. 10 Ошибки второго рода при поиске шестеренки (4, 20, 70, 60, 45)

белый фон) — ошибка первого рода. В то же время, если позволить считать успешным распознаванием слишком слабое совпадение, то в результат войдут области, не имеющие практически ничего общего с образцом (рис. 10) — ошибка второго рода.

Ранее отмечалось, что управляющие параметры следует подбирать исходя из желаемой «грубости» поиска. Приведем некоторые формулы, позволяющие выбрать значения этих параметров в условиях конкретной задачи.

В зависимости от параметра ширины спектра при переходе от ИПР к ее ФК можно менять количество Фурье-резонаторов, используемых при преобразовании. Их число можно получить по формуле spectrumWidth · projection.length. Параметр вычисляемых первых ФК для ИПР следует умножить на это число, тогда получим количество ФК для каждой ИПР в ПФД изображения: considerationPercent · spectrumWidth · projection.length.

Общая формула для минимального размера учитываемого элемента изображения будет выглядеть так: min {img.height, img.width}/(considerationPercent · spectrumWidth).

3.2 Дорожные знаки

Теперь отойдем от тестовых изображений и покажем эффективность работы алгоритма при решении реальных задач.

Например, существует задача распознавания дорожных знаков, качественного решения которой на данный момент не существует. Описанный алгоритм может стать таким решением. Возьмем в качестве образца изображение знака «пешеходный переход», четкое и ровное (рис. 11).

Рис. 11 Изображение знака «пешеходный переход»

В качестве изображения теста возьмем фотографию с реальным дорожным знаком (рис. 12).

Рис. 12 Реальный дорожный знак

Выделим в этом изображении сам квадрат знака (рис. 13) (например, с помощью примитивов Хаара [7]) и попробуем распознать в нем образец.

При точном выборе управляющих параметров поиска удается распознать на фотографии «образцовый» дорожный знак (рис. 14).

Рис. 13 Выделенный из фотографии квадрат знака

Рис. 14 Результат распознавания знака (10, 7, 75, 70, 70)

Отметим, что в результатах отображаются все найденные совпадения, и они являются очень точными. Это значит, что удалось избежать ошибок второго рода. Также удалось избежать и ошибок первого рода: можно видеть, что знак распознан верно, несмотря на дефект при печати самого знака (горизонтальная белая линия в центре) и надпись внизу знака, искажающую его часть.

3.3 Отслеживание глаз (айтрекинг)

Другая практическая задача, нуждающаяся в решении, — качественное распознавание зрачка человека. Существующие методы не дают высокой точности, и в результате при попытке отслеживать движения зрачка в видеопотоке происходит «дрожание» распознанного зрачка, даже если в реальности зрачок неподвижен. Рассматриваемый в данной статье алгоритм распознавания может решить эту проблему. Продемонстрируем это на примере. Единожды указав в качестве образца зрачок с частью радужной оболочки (рис. 15), можно отслеживать его положение в кадре.

Рис. 15 Изображение зрачка глаза человека и части радужной оболочки

Теперь попробуем найти зрачки в двух разных кадрах видео (рис. 16).

Выделим в этих кадрах области глаз (рис. 17).

После этого запустим поиск зрачков в полученных изображениях. Очевидно, что программа справляется с этой задачей, причем с минимальными ошибками второго рода, которые, в свою очередь, можно устранить выбором наиболее близкой к образцу области в районе каждого глаза. Снимки экрана с результатами представлены на рис. 18 и 19.

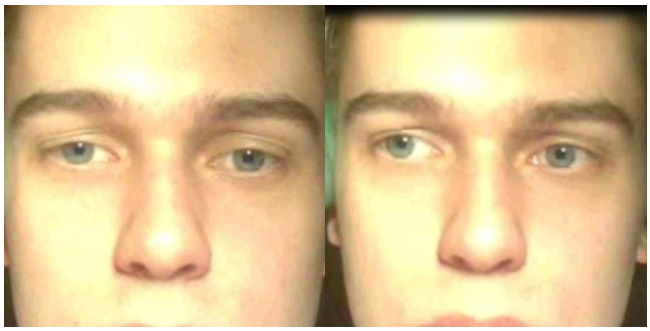


Рис. 16 Два различных кадра из видеозаписи с лицом человека

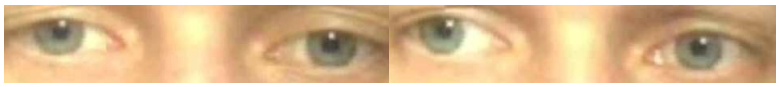


Рис. 17 Выделенные из кадров фрагменты с изображениями глаз человека

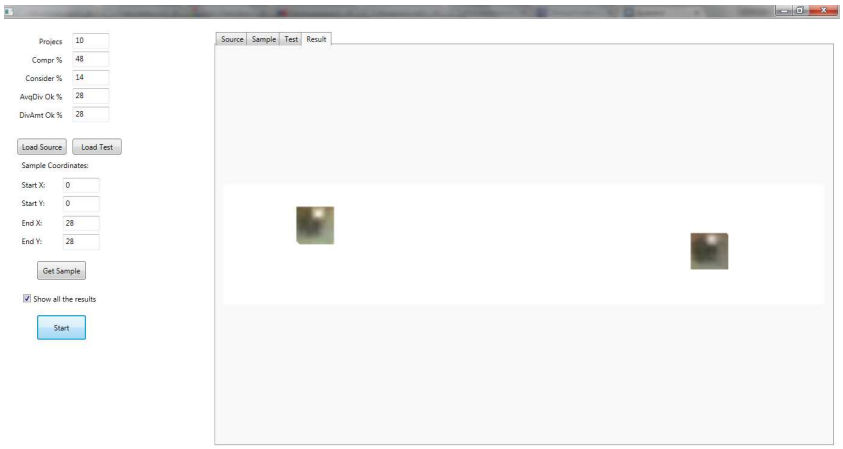


Рис. 18 Результат поиска зрачков в фрагменте из первого кадра (10, 48, 14, 28, 28)

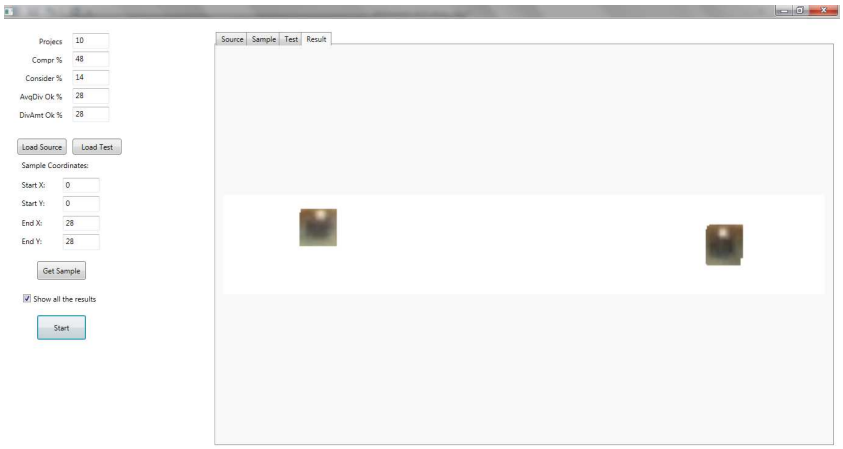


Рис. 19 Результат поиска зрачков в фрагменте из второго кадра (10, 48, 14, 28, 28)

3.4 Перспективы

За счет универсальности подхода при анализе изображения рассмотренный метод может применяться в широком спектре задач, возникающих в области обработки данных.

В современном мире в условиях постоянно растущего объема передаваемых по различным каналам данных задача их обработки становится одной из наиболее актуальных [8]. Описанный метод позволяет приблизиться к решению этой задачи в случае, когда в качестве передаваемых данных служит информация о свойствах изображения и его содержании. Анализируя ПФД некоторого изображения, можно получить эту информацию, причем объем данных, занимаемых ПФД, значительно меньше, чем объем сильно сжатого (даже с потерей данных) изображения.

Другая сфера применения метода — контроль качества графических данных. Как упоминалось ранее, используемый алгоритм способен бороться с шумами и помехами в изображениях. Этот механизм также позволяет анализировать количество и интенсивность шумов и помех, позволяя тем самым делать выводы о качестве изображений или видеопотоков.

Еще одна задача обработки данных, одна из самых сложных и вместе с тем актуальных, — задача распознавания. Она относится к области «машинного зрения» (Computer Vision, или CV). На сегодняшний день существует множество различных методов анализа объектов в изображении, однако все они являются сильно специфическими и приспособлены для решения задач в узких областях [9]. Представленный метод отличает универсальность, т. е. он может применяться для решения задач распознавания любых объектов с любой заранее выбранной глубиной точности. Также этот метод может применяться при отслеживании положения объектов в кадре, т. е. в задаче ведения распознанного объекта в видеопотоке. Такая необходимость возникает, например, в системах автоматического мониторинга видеонаблюдения, мониторинга контента телевизионных каналов, отслеживания движений человека. В частности, метод позволяет с высокой точностью отслеживать движения зрачков человека, что находит применение в компьютерной окулографии.

Для повышения толерантности алгоритма к поворотам сравниваемых матриц изображений относительно друг друга необходимо использовать дополнительный механизм определения начальной проекции для сравнения. При этом достаточно использовать только первые несколько ФК для подряд идущих проекций. Это позволит добиться распознавания искомого объекта с произвольным углом поворота при незначительном увеличении вычислительной нагрузки.

Чтобы справиться с задачей распознавания, когда разрешение изображения искомого объекта отличается от разрешения изображения-образца, достаточно добиться синхронизации количества результирующих ФК в ПФД тестируемого изображения и количества ФК в ПФД изображения-образца. Для этого необходимо умножить значение исходной ширины спектра на коэффициент $k = \text{size}_{\text{sample}}/\text{size}_{\text{test}}$ отношения размера изображения-образца к размеру области-теста $\text{spectrumWidth}_{\text{test}} = k \cdot \text{spectrumWidth}_{\text{sample}}$ и производить преобразование Фурье вдоль проекции изображения-теста с шагом, кратным тому же коэффициенту k. Совпадение значений ФК для одного объекта на изображениях разных размеров наглядно видно на графиках, изображенных на рис. 20.

В перспективе с использованием механизмов «машинного обучения» (Machine Learning, или ML) описанный метод может применяться для создания системы интеллектуального анализа графических данных. При этом масштабы задействованных вычислительных мощностей и объемы передаваемых данных будут несравнимо меньше, чем у подобных систем, основанных на искусственных нейронных сетях.

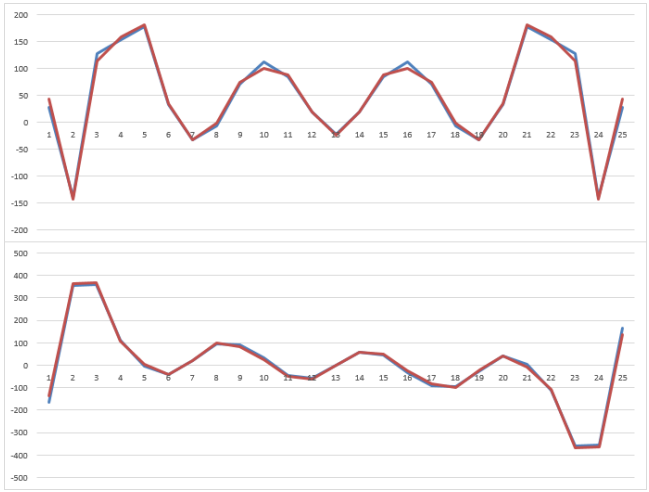


Рис. 20 Графики значений ФК для изображений разных размеров

4 Заключение

Представленный в данной статье метод, основанный на комбинации классических преобразований Радона и Фурье, а также статистическом анализе, реализован в виде рабочего алгоритма, успешно решающего поставленную задачу мониторинга видеопотока и анализа изображений. Метод уже внедрен для решения некоторых задач, в частности в системе компьютерной окулографии, используемой для диагностики нейропатологий через отслеживание движений глаз (айтрекинг).

Получены данные о сильных и слабых сторонах алгоритма и возможных методах оптимизации исполняемого кода программного прототипа.

С учетом свойств данных, полученных при применении метода к конкретному изображению, были сделаны выводы о возможных областях применения алгоритма и перспективах его использования.

Литература

[1] Новая информационная технология обработки произвольных изображений — ТАРе-технология. `http://comexp.ru/node/129`.

[2] *Грузман И. С.* Математические задачи компьютерной томографии // Соросовский образовательный журнал, 2001. № 5.

[3] *Зорич В. А.* Математический анализ. — М.: Физматлит, 1984. 544 с.

[4] *Тихонов А. Н., Арсенин В. Я., Тимонов А. А.* Математические задачи компьютерной томографии. Серия: Проблемы науки и технического прогресса. — М.: Наука, 1987. 160 с.

[5] *Gertner I.* New efficient algorithm to complete the two-dimensional discrete Fourier transform // IEEE Trans. ASSP, 1988. Vol. 36. No. 7. P. 1036–1050.

[6] *Сергиенко А. Б.* Цифровая обработка сигналов. — 2-е изд. — СПб.: Питер, 2006. 751 с.

[7] *Lienhart R., Maydt J.* An extended set of Haar-like features for rapid object detection // ICIP02, 2002. P. 900–903.

[8] *Горелик А. Л., Скрипкин В. А.* Методы распознавания. — 4-е изд. — М.: Высшая школа, 2004. 262 с.

[9] *Фомин Я. А.* Распознавание образов: теория и применения. — 2-е изд. — М.: ФАЗИС, 2012. 429 с.

Поступила в редакцию 23.07.2015

References

[1] Novaya informatsionnaya tekhnologiya obrabotki proizvol'nykh izobrazheniy — TAPe-tekhnologiya. [New information technology of processing arbitrary images — TAPe-technology]. Available at: `http://comexp.ru/node/129` (accessed December 11, 2015). (In Russian.)

[2] Gruzman, I. S. 2001. Matematicheskie zadachi komp'yuternoy tomografii [Mathematical problems of computed tomography]. *Sorosovskiy Obrazovatel'nyy Zhurnal* 5. (In Russian.)

[3] Zorich, V. A. 1984. *Matematicheskiy analiz* [Mathematical analysis]. Moscow: Fizmatlit. 544 p. (In Russian.)

[4] Tikhonov, A. N., V. Ya. Arsenin, and A. A. Timonov. 1987. Matematicheskie zadachi komp'yuternoy tomografii [Mathematical problems of computed tomography]. Ser. Problemy nauki i tekhnicheskogo progressa [Problems of science and technological progress ser.] Moscow: Nauka. 160 p. (In Russian.)

[5] Gertner, I. 1988. New efficient algorithm to complete the two-dimensional discrete Fourier transform. *IEEE Trans. ASSP* 36(7):1036–1050.

[6] Sergienko, A. B. 2006. *Tsifrovaya obrabotka signalov* [Digital signal processing]. 2nd ed. St. Petersburg: Piter. 751 p. (In Russian.)

[7] Lienhart, R., and J. Maydt. 2002. An extended set of Haar-like features for rapid object detection. *ICIP02*. 900–903.

[8] Gorelik, A. L., and V. A. Skripkin. 2004. *Metody raspoznavaniya* [Methods of recognition]. 4th ed. Moscow: Vysshaya Shkola. 262 p. (In Russian.)

[9] Fomin, Ya. A. 2012. *Raspoznavanie obrazov: Teoriya i primeneniya* [Pattern recognition: Theory and application]. 2nd ed. Moscow: FAZIS. 429 p. (In Russian.)

Received July 23, 2015

Комбинированная нелинейная фильтрация цифровых изображений большой разрядности

Е. П. Петров, Н. Л. Харина, Е. Д. Ржаникова
eppetrov@mail.ru
ФГБОУ ВО «Вятский государственный университет», г. Киров

Синтезирован алгоритм нелинейной фильтрации многоразрядных цифровых изображений (ЦИ), передаваемых многопозиционными фазоманипулированными (ФМ) импульсными сигналами, что позволяет сократить время передачи ЦИ. Синтезированный алгоритм реализует пространственную и межразрядную статистическую избыточность многоразрядных ЦИ для компенсации потерь помехоустойчивости при переходе от двухпозиционных ФМ сигналов к многопозиционным ФМ сигналам. В комбинации с медианной фильтрацией алгоритм нелинейной фильтрации многоразрядных ЦИ может подавлять не только белый гауссовский шум (БГШ), но и импульсные помехи, борьба с которыми медианной фильтрацией при наличии БГШ неэффективна.

Ключевые слова: *цифровое изображение; нелинейная фильтрация; цепь Маркова*

DOI: 10.21469/22233792.1.13.06

Combined nonlinear filtration of digital halftone high bitness images

E. P. Petrov, N. L. Kharina, and E. D. Rzhanikova
Vyatka State University, 36 Moskovskaya st., Kirov, Russia

The requirement for transfer of a large volume of information, such as a multibit digital images (DI), more quickly is an actual task and demands perfecting of radiocommunication means. One of the ways of reduction of a DI transfer time is transition to a multiphase frequency modulation (FM) signals. However, their application is limited because of a noise stability loss at each division of a phase in comparison with binary FM signals. At the transfer of DI by the eight-phase FM signals, the time is reduced by four times, but with partial compensation of a noise stability loss. The algorithm of restoration of a multibit DI distorted by white Gaussian noise (WGN) is developed. The statistical redundance of the DI is efficiently used for compensation of a noise stability loss at the transfer of digital images by multiphase FM signals. For example, the time of the DI transfer by four-phase signals was reduced twice without noise stability loss in comparison with the DI transfer by the binary FM signals. The combined algorithm of filtration of multidigit DI is constructed. It consists of two algorithms: a nonlinear filtration of DI distorted by WGN and the median filter for restoration of DI distorted by salt–pepper impulse noise. Due to separation of impulse noise and WGN, the impulse noise is efficiently suppressed by the median filter. The results of such combination allow to reduce transfer time of a multibit DI and to strive successfully against WGN and impulse noise.

Keywords: *digital image; nonlinear filtering; Markov chain*

DOI: 10.21469/22233792.1.13.06

1 Введение

При прямой передаче многоразрядных ЦИ по каналу связи с помехами, например, БГШ и импульсными помехами типа «перец-соль», требуются большие временны́е и энергетические ресурсы. Сократить первое и второе можно, если для передачи ЦИ использовать многофазные ФМ (МФМ) сигналы, а потери помехоустойчивости, возникающие при этом, скомпенсировать полностью или частично реализацией статистической избыточности ЦИ.

Для решения этой задачи необходимо синтезировать алгоритм нелинейной фильтрации ЦИ в присутствии БГШ с импульсными помехами, которые могут быть подавлены, например, медианным фильтром, подключенным к выходу синтезированного нелинейного фильтра, т. е. необходимо разработать алгоритм комбинированной нелинейной фильтрации ЦИ при наличии БГШ и импульсных помех, в котором неизвестным является алгоритм нелинейной фильтрации ЦИ, передаваемого МФМ импульсными сигналами, основным показателем которого является эффективная реализация статистической избыточности, для повышения помехоустойчивости приема ЦИ.

Будем полагать, что g-разрядное ($g \geqslant 8$) ЦИ является марковским случайным полем (МСП) с 2^g дискретными состояниями (градациями яркости). Для синтеза алгоритма фильтрации ЦИ необходимо построить математическую модель (ММ) многоразрядного ЦИ.

2 Постановка задачи

Необходимо разработать алгоритм нелинейной фильтрации ЦИ, эффективно использующий статистическую избыточность ЦИ для повышения качества восстановления ЦИ, искаженных БГШ $n(t)$ с нулевым средним и дисперсией σ_n^2.

3 Математическая модель многоразрядного цифрового изображения

Будем полагать, что g-разрядное ЦИ состоит из g разрядных двоичных изображений (РДИ), каждое из которых МСП — двумерная цепь Маркова с двумя равновероятными ($p_1 = p_2$) состояниями (рис. 1).

Для решения задачи сокращения времени передачи многоразрядных ЦИ объединим в g-разрядном ЦИ соседние РДИ в равные группы. Например, в 16-разрядном ЦИ можно образовать группы по 2 (рис. 2, *а*) или 4 соседних РДИ. На рис. 2, *б* представлена группа из двух старших РДИ 16-разрядного ЦИ, в каждом столбце которой два бинарных пикселя могут принимать четыре равновероятных ($p_1 = p_2 = p_3 = p_4$) состояния. В результате объединения двух РДИ получаем групповое разрядное ЦИ (ГРЦИ) с четырьмя градациями яркости.

Передачу ГРЦИ можно осуществлять четырехфазными импульсными сигналами. Схема образования ГРЦИ и переход от двоичной фазовой манипуляции к квадратурной показан на рис. 3. Применение МФМ сигналов для передачи ГРЦИ вместо РДИ позволяет сократить время передачи ЦИ в число раз, равное числу РДИ в ГРЦИ, поскольку за одну единицу времени по радиоканалу передается не один бит информации, как при двоичной фазовой манипуляции, а несколько, например два в случае объединения двух РДИ.

Если РДИ — двумерная цепь Маркова с двумя состояниями [1], то будем полагать ГРЦИ двумерной цепью Маркова с вектором вероятностей из N начальных состояний

$$P = \left\| p_1, \quad p_2, \quad \ldots, \quad p_n \right\|^{\mathrm{T}} \tag{1}$$

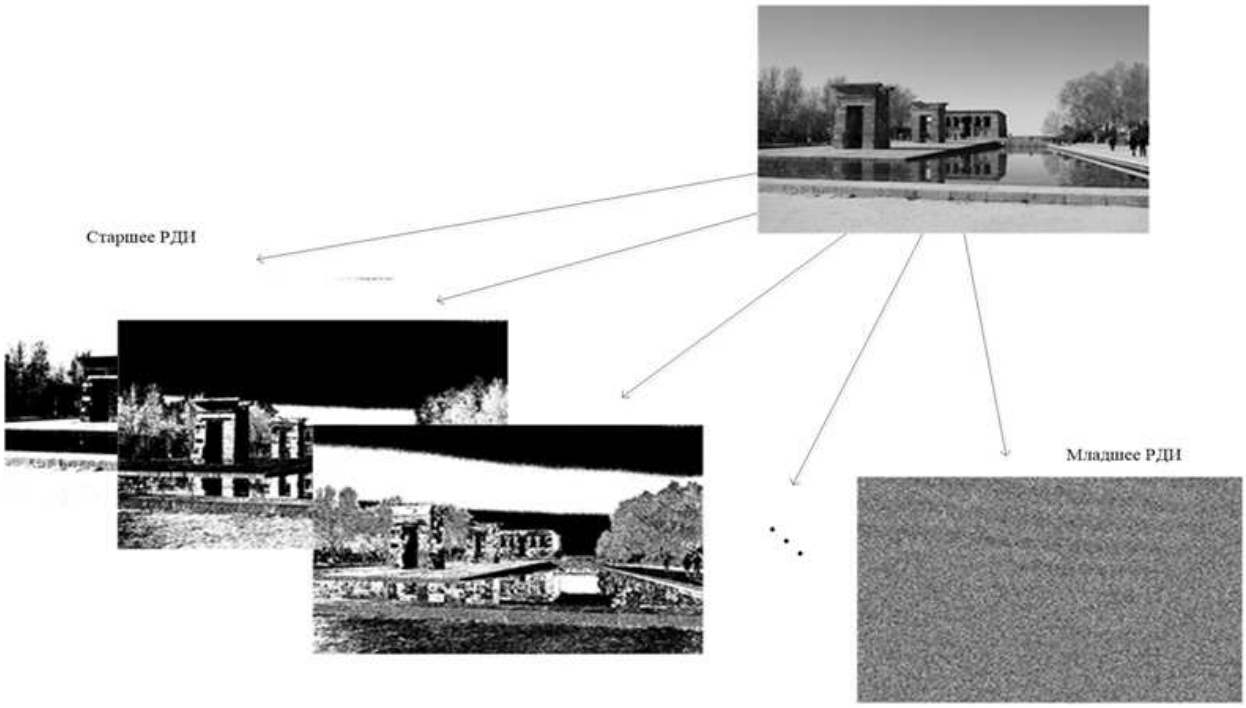
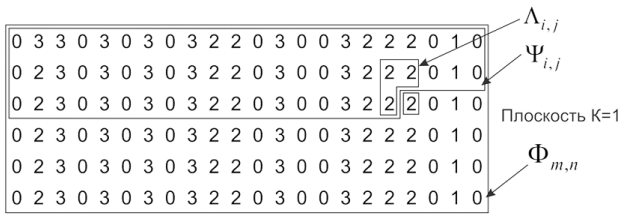


Рис. 1 Представление ЦИ набором независимых РДИ

(*а*) Разрядное двоичное изображение — 15 и 16 разрядов



(*б*) Групповое разрядное ЦИ, полученное объединением РДИ

Рис. 2 Попарное объединение разрядных плоскостей

и матрицами вероятностей перехода (МВП) из i-го состояния в j-е за один шаг:

$$^1\Pi = \|^1\pi_{ij}\|_{N\times N}; \quad ^2\Pi = \|^2\pi_{ij}\|_{N\times N}, \quad i \neq j. \tag{2}$$

Элементы МВП (2) удовлетворяют условию нормировки

$$\sum_{j=1}^{N} {}^q\pi_{ij} = 1, \quad i = \overline{1,N}, \quad q = \overline{1,2},$$

и стационарности

$$p_i = \sum_{j=1}^{N} p_j\pi_{ij}, \quad i = \overline{1,N}.$$

На рис. 4, *а* приведена модель ГРЦИ [2], удовлетворяющая априорно заданным: вектору (1) и МВП (2). Реализация модели и ее адекватность реальным ЦИ подробно исследована в работах [1, 3]. Размер окрестности элемента M_{ij} можно взять произвольным,

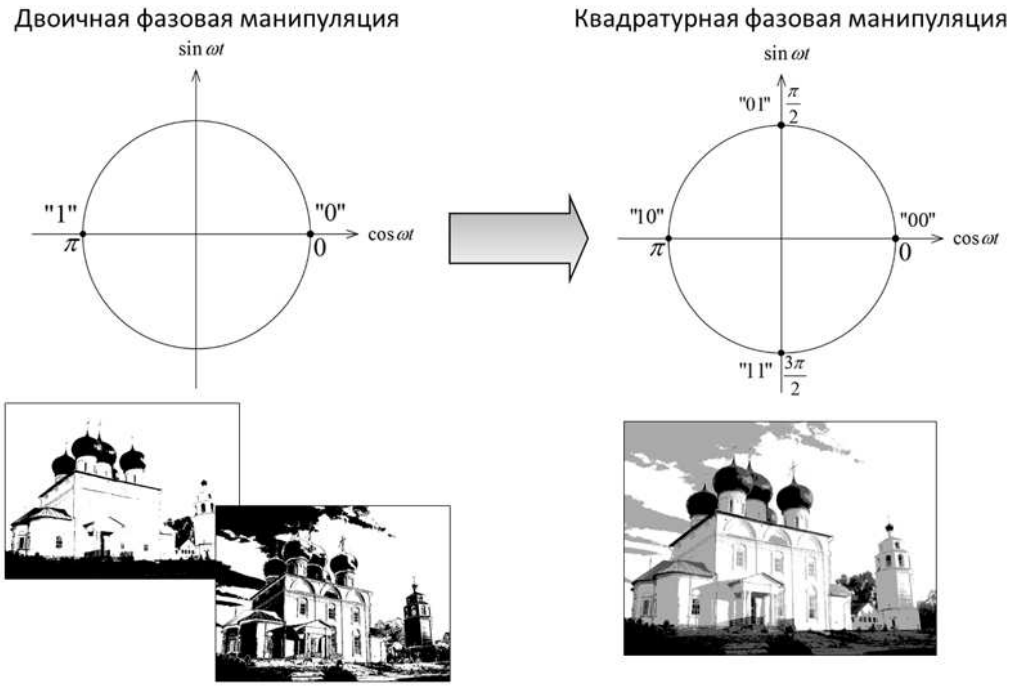
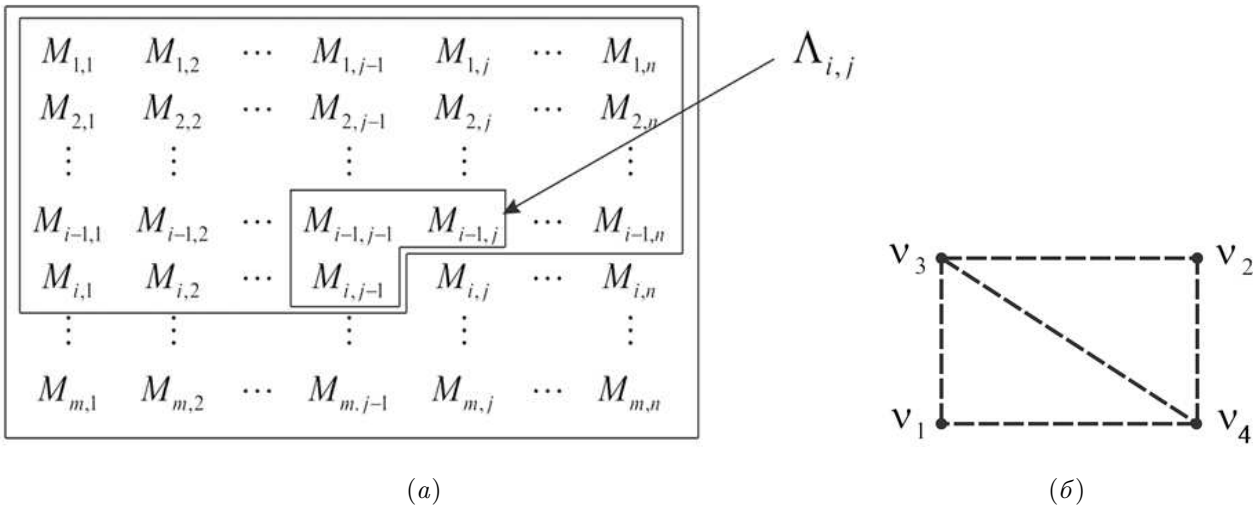


Рис. 3 Переход от двоичной к квадратурной фазовой манипуляции



(*а*) (*б*)

Рис. 4 Математическая модель ЦИ (*а*) и окрестность элемента изображения (*б*)

но ее увеличение приводит к сложным двумерным цепям Маркова [2] и практически мало улучшает качество искаженного шумом ЦИ [4], поэтому выбираем окрестность вида рис. 4, *б*.

Предположим, что в 16-разрядном ЦИ ГРЦИ состоит из двух РДИ и каждый фильтруемый элемент $M_{i,j}$ ГРЦИ (см. рис. 4, *а*) зависит только от соседних ранее известных элементов ГРЦИ, образующих окрестность $\Lambda_{i,j}$ элемента ν_4, где приняты обозначения: $\nu_1 = M_{i,j-1}$, $\nu_2 = M_{i-1,j}$, $\nu_3 = M_{i-1,j-1}$ [2].

Для ГРЦИ из двух РДИ вероятности перехода от комбинаций состояний элементов окрестности $\Lambda_{i,j}$ к элементу ν_4 (см. рис. 4, *б*) образуют МВП вида:

$$\Pi = \left\| \begin{matrix} \pi_{iii} & \pi_{iji} & \pi_{iki} & \pi_{ili} & \pi_{jii} & \pi_{jji} & \pi_{jki} & \pi_{jli} & \cdots & \pi_{lii} & \pi_{lji} & \pi_{lki} & \pi_{lli} \\ \pi_{iij} & \pi_{ijj} & \pi_{ikj} & \pi_{ilj} & \pi_{jij} & \pi_{jjj} & \pi_{jkj} & \pi_{jlj} & \cdots & \pi_{lij} & \pi_{ljj} & \pi_{lkj} & \pi_{llj} \\ \pi_{iik} & \pi_{ijk} & \pi_{ikk} & \pi_{ilk} & \pi_{jik} & \pi_{jjk} & \pi_{jkk} & \pi_{jlk} & \cdots & \pi_{lik} & \pi_{ljk} & \pi_{lkk} & \pi_{llk} \\ \pi_{iil} & \pi_{ijl} & \pi_{ikl} & \pi_{ill} & \pi_{jil} & \pi_{jjl} & \pi_{jkl} & \pi_{jll} & \cdots & \pi_{lil} & \pi_{ljl} & \pi_{lkl} & \pi_{lll} \end{matrix} \right\|. \qquad (3)$$

Элементы первого столбца МВП Π (3) связаны с элементами матриц (2) следующими соотношениями (остальные вычисляются аналогично):

$$\pi_{iii} = \frac{{}^1\pi_{ii}\,{}^2\pi_{ii}}{{}^3\pi_{ii}}; \quad \pi_{iij} = \frac{{}^1\pi_{ij}\,{}^2\pi_{ij}}{{}^3\pi_{ii}}; \quad \pi_{iik} = \frac{{}^1\pi_{ik}\,{}^2\pi_{ik}}{{}^3\pi_{ii}}; \quad \pi_{iil} = \frac{{}^1\pi_{il}\,{}^2\pi_{il}}{{}^3\pi_{ii}};$$
$$\pi_{iji} = \frac{{}^1\pi_{ii}\,{}^2\pi_{ji}}{{}^3\pi_{ij}}; \quad \pi_{ijj} = \frac{{}^1\pi_{ij}\,{}^2\pi_{jj}}{{}^3\pi_{ij}}; \quad \pi_{ijk} = \frac{{}^1\pi_{ik}\,{}^2\pi_{jk}}{{}^3\pi_{ij}}; \quad \pi_{ijl} = \frac{{}^1\pi_{il}\,{}^2\pi_{jl}}{{}^3\pi_{ij}},$$

где ${}^3\pi_{ii}$ — элементы дополнительной матрицы ${}^3\Pi = {}^1\Pi \times {}^2\Pi^{\mathrm{T}}$, связывающей ν_3 с ν_4.

4 Синтез алгоритма нелинейной фильтрации многоразрядных цифровых изображений

Пусть ГРЦИ многоразрядного ЦИ передаются по каналу связи четырехфазными ФМ сигналами при наличии БГШ $n(t)$ с нулевым средним и дисперсией σ_n^2.

Используя теорию фильтрации условных марковских процессов с дискретными аргументами [5], синтезируем алгоритмы нелинейной фильтрации многоразрядных ЦИ, представленных ГРЦИ.

Опуская процедуру синтеза алгоритма нелинейной фильтрации ГРЦИ, которая аналогична процедуре синтеза алгоритмов нелинейной фильтрации РДИ [1, 3], запишем систему рекуррентных уравнений нелинейной фильтрации ГРЦИ, представляющего двумерную цепь Маркова с четырьмя состояниями в виде [4, 7]:

$$\begin{aligned} u_1(\nu_4) = {} & [f(M_1(\nu_4)) - f(M_4(\nu_4))] + u_1(\nu_1) + z_1\left(u(\nu_1), {}^1\pi_{ij}\right) + \\ & + u_1(\nu_2) + z_1\left(u(\nu_2), {}^2\pi_{ij}\right) - u_1(\nu_3) - z_1\left(u(\nu_3), {}^3\pi_{ij}\right); \\ u_2(\nu_4) = {} & [f(M_2(\nu_4)) - f(M_4(\nu_4))] + u_2(\nu_1) + z_2\left(u(\nu_1), {}^1\pi_{ij}\right) + \\ & + u_2(\nu_2) + z_2\left(u(\nu_2), {}^2\pi_{ij}\right) - u_2(\nu_3) - z_2\left(u(\nu_3), {}^3\pi_{ij}\right); \\ u_3(\nu_4) = {} & [f(M_3(\nu_4)) - f(M_4(\nu_4))] + u_3(\nu_1) + z_3\left(u(\nu_1), {}^1\pi_{ij}\right) + \\ & + u_3(\nu_2) + z_3\left(u(\nu_2), {}^2\pi_{ij}\right) - u_3(\nu_3) - z_3\left(u(\nu_3), {}^3\pi_{ij}\right), \end{aligned}$$

где $u_j(\nu_4) = \ln[p_j(\nu_4)/p_4(\nu_4)]$ — апостериорная вероятность дискретного параметра МФМ импульсных сигналов, адекватных состояниям элементов ГРЦИ; $\left[f(M_i(\nu_4)) - f(M_4(\nu_4))\right]$, $i = \overline{1,3}$, — разность логарифмов функции правдоподобия состояний дискретного параметра МФМ импульсных сигналов (элементов ГРЦИ); $z_j(\cdot)$ — нелинейная функция вида:

$$z_j\left(u(\nu_l), {}^l\pi_{ij}\right) = \ln\left[\frac{\sum_{i=1, i\neq j}^{3}\left\{\exp\left(u_i(\nu_l) - u_j(\nu_l)\right){}^l\pi_{ij} + \exp\left(-u_j(\nu_l)\right){}^l\pi_{ij} + \pi_{jj}\right\}}{\sum_{i=1}^{3}\left\{\exp\left(u_j(\nu_l)\right){}^l\pi_{i4}\right\} + \pi_{44}}\right] \left(j = \overline{1,3}, l = \overline{1,3}\right). \qquad (4)$$

Вся априорная информация о статистической зависимости состояний элементов ГРЦИ сосредоточена в слагаемых вида (4), где ${}^{l}\pi_{ij}$, $\left(i,j=\overline{1,4},l=\overline{1,3}\right)$ — элементы матриц ${}^{1}\Pi$, ${}^{2}\Pi$ и ${}^{3}\Pi$.

В качестве критерия различения состояний элементов ГРЦИ принят критерий максимума логарифма отношения апостериорных вероятностей $u_j(\nu_4)\left(j=\overline{1,3}\right)$, в соответствии с которым, если

$$u_j(\nu_4) > u_i(\nu_4)\,, \quad i,j=\overline{1,3}\,,\ \ i\neq j\,,$$

то принимается решение о состоянии элемента изображения $\nu_4 = M_j$, если все значения $u_j(\nu_4) < 0\ \left(j=\overline{1,3}\right)$, то принимается решение о состоянии элемента изображения $\nu_4 = M_4$.

5 Комбинированный алгоритм фильтрации многоразрядных цифровых изображений

Для борьбы с импульсными помехами часто применяют медианную фильтрацию, которая при наличии БГШ неэффективна, а параметрическая нелинейная фильтрация не подавляет импульсные помехи, а, напротив, выделяет их из БГШ как мелкие объекты в многоразрядных ЦИ. Поэтому целесообразным является разделение функций, выполняемых и той, и другой нелинейными фильтрациями, т. е. вначале фильтруем ЦИ, состоящее из ГРЦИ, при наличии БГШ, а затем к выделенным из шума импульсным помехам применяем медианную фильтрацию. При этом алгоритм фильтрации приобретает комбинированный характер, обеспечивая эффективное подавление обеих помех.

В совокупности с медианным фильтром разработанный нелинейный фильтр образует комбинированный нелинейный фильтр (рис. 5), который позволяет успешно бороться с БГШ и импульсными помехами [7].

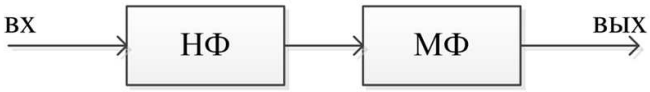


Рис. 5 Схема комбинированного нелинейного фильтра

6 Результаты исследования

Анализ результатов нелинейной фильтрации показывает, что снижение помехоустойчивости приема ЦИ, вызванное переходом к МФМ импульсным сигналам, удается полностью скомпенсировать применением разработанного алгоритма за счет использования статистической избыточности, содержащейся в цифровых изображениях.

Ниже приведен пример работы комбинированного нелинейного фильтра. На рис. 6 представлено исходное тестовое 16-разрядное ЦИ. На рис. 7, *а* представлен фрагмент тестового изображения, передаваемого четырехфазными сигналами и искаженного БГШ при отношении сигнал/шум по мощности $\rho^2 = -6$ дБ.

На рис. 7, *в* показан фрагмент ЦИ при передаче четырехфазными сигналами на выходе нелинейного фильтра, а на рис. 7, *д* — на выходе комбинированного фильтра. На рис. 7, *б*, 7, *г* и 7, *е* приведены результаты нелинейной и комбинированной фильтрации ЦИ, передаваемого бинарными сигналами. Для оценки качества фильтрации были вычислены значения среднеквадратической ошибки (СКО) ЦИ на входе и выходе нелинейного и комбинированного фильтров. В результате фильтрации ЦИ, передаваемого бинарными

Рис. 6 Исходное изображение

сигналами, значение СКО уменьшилось в 7–8 раз, при фильтрации ЦИ, передаваемого четырехфазными сигналами, — в 12 раз.

7 Заключение

Переход от ЦИ к ГРЦИ, передаваемых четырехкратными ФМ сигналами, позволил сократить время передачи 16-разрядного ЦИ в 2 раза и полностью скомпенсировать за счет реализации статистической избыточности ГРЦИ бóльшую часть потерь в помехоустойчивости, вызванных переходом от двухкратных ФМ сигналов для передачи РДИ в ЦПИ к четырехкратным ФМ сигналам для передачи ЦИ из ГРЦИ.

Литература

[1] *Петров Е. П., Трубин И. С., Частиков И. А.* Нелинейная фильтрация видеопоследовательностей цифровых полутоновых изображений марковского типа // Успехи современной радиоэлектроники, 2007. № 3. С. 54–87.

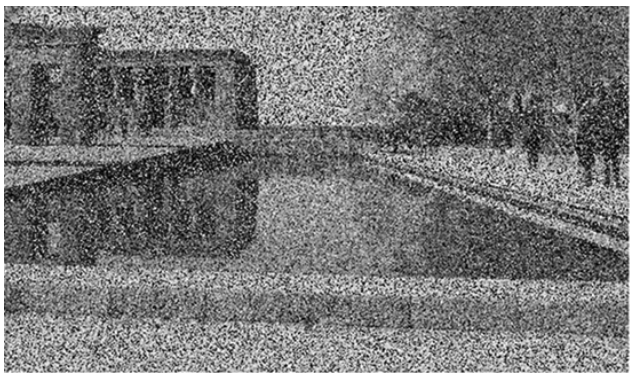
[2] *Петров Е. П., Харина Н. Л., Ржаникова Е. Д.* Математическая модель цифровых полутоновых изображений на основе цепей Маркова с несколькими состояниями // Нелинейный мир, 2013. Т. 11. № 7. С. 487–492.

[3] *Петров Е. П., Харина Н. Л., Ржаникова Е. Д.* Синтез и исследование алгоритмов фильтрации дискретных марковских процессов с несколькими состояниями // Радиотехнические и телекоммуникационные системы, 2013. № 1. С. 60–66.

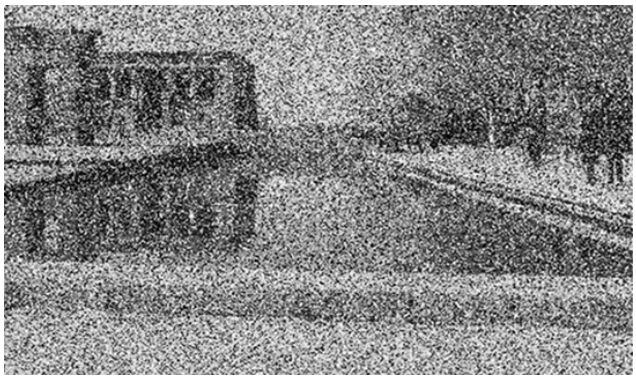
[4] *Петров Е. П., Харина Н. Л., Харюшин В. Ф.* Математические модели и алгоритмы фильтрации цифровых полутоновых изображений на основе сложных цепей Маркова // Цифровая обработка сигналов, 2012. № 3. С. 52–57.

[5] *Амиантов И. Н.* Избранные вопросы статистической теории связи. — М: Советское радио, 1971. 416 с.

[6] *Петров Е. П., Харина Н. Л., Ржаникова Е. Д.* Нелинейная фильтрация изображений на основе цепей Маркова с несколькими состояниями // Мат-лы III Всеросс. НТК «Актуальные проблемы ракетно-космической техники». — Самара, 2013. С. 154–163.

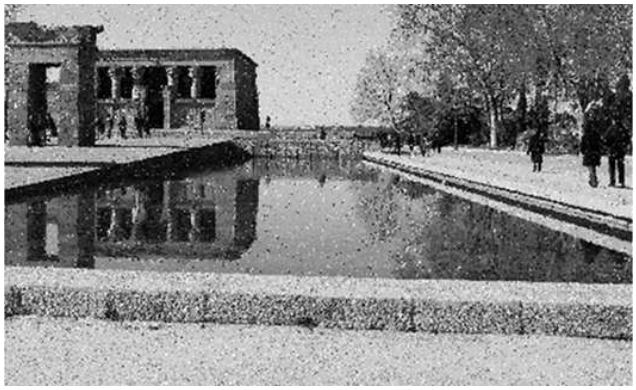


(*а*) Зашумленное ЦИ, передаваемое четырехфазными сигналами (−6 дБ), СКО = $2{,}84 \cdot 10^8$



(*б*) Зашумленное ЦИ, передаваемое бинарными сигналами (−6 дБ), СКО = $3{,}09 \cdot 10^8$

(*в*) Восстановленное ЦИ, передаваемое четырехфазными сигналами, СКО = $0{,}22 \cdot 10^8$



(*г*) Восстановленное ЦИ, передаваемое бинарными сигналами, СКО = $0{,}39 \cdot 10^8$

(*д*) Цифровое изображение, передаваемое четырехфазными сигналами, на выходе комбинированного фильтра, СКО = $0{,}11 \cdot 10^8$

(*е*) Цифровое изображение, передаваемое бинарными сигналами, на выходе комбинированного фильтра, СКО = $0{,}10 \cdot 10^8$

Рис. 7 Пример работы комбинированного фильтра

[7] *Медведева Е. В., Метелев А. П.* Метод комбинированной нелинейной фильтрации коррелированных видеоизображений // Нелинейный мир, 2010. № 11. С. 677–684.

Поступила в редакцию 15.06.2015

References

[1] Petrov, E. P., I. S. Trubin, and I. A. Chastikov. 2007. Non-linear filtration of video sequences of digital halftone images of Markov type. *Uspekhi sovremennoi radioelektroniki* 3:54–87.

[2] Petrov, E. P., N. L. Kharina, and E. D. Rzhanikova. 2013. Mathematical model of digital halftone images based on Markov chains with several states. *Nonlinear World* 11(7):487–492.

[3] Petrov, E. P., N. L. Kharina, and E. D. Rzhanikova. 2013. Synthesis and analysis of algorithms for filtering of discrete Markov processes with several states. *Radiotechnical Telecommunication Syst.* 1:60–66.

[4] Petrov, E. P., N. L. Kharina, and V. F. Kharyushin. 2012. Mathematical models and filtration algorithms of digital halftone images based on complex Markov chains. *Digital Signal Processing* 3:52–57.

[5] Amiantov, I. N. *Selected questions of the statistical theory of communication*. Moscow: Sovetskoe Radio. 314 p.

[6] Petrov, E. P., N. L. Kharina, and E. D. Rzhanikova. 2013. Non-linear image filtration based on Markov chains with several states. *3rd All-Russian Scientific and Technical Conference "Actual Problems of the Missile and Space Equipment" Proceedings*. Samara. 154–163.

[7] Medvedeva, E. V., and A. P. Metelev. 2010. Method of combined non-linear filtration of correlated video images. *Nonlinear World* 11:677–684.

Received June 15, 2015

Об эффективном распараллеливании алгоритмов для дискретных перечислительных задач*

***Е. В. Дюкова*[1], *А. Г. Никифоров*[2]**
edjukova@mail.ru, ankifor@gmail.com
[1]Вычислительный центр РАН им. А. А. Дородницына, Москва, Россия
[2]Московский государственный университет им. М. В. Ломоносова, Москва, Россия

Разработана новая статическая схема распараллеливания асимптотически оптимальных алгоритмов для задачи дуализации. Данная задача относится к числу труднорешаемых перечислительных задач. Предлагаемая схема основана на предварительной статистической обработке входных данных с целью установления вида распределения случайной величины, определяющей объемы подзадач. Статья является развитием ранней работы авторов, в которой при получении указанных оценок использовалась менее эффективная методика, учитывающая только размер задачи. Выявлены условия, при которых обеспечиваются достаточно равномерная загрузка процессоров и ускорение, близкое к максимальному.

Ключевые слова: *перечислительная задача; дуализация; неприводимое покрытие булевой матрицы; трансверсаль гиперграфа; асимптотически оптимальный алгоритм; параллельные вычисления; балансировка нагрузки; сильная масштабируемость*

DOI: 10.21469/22233792.1.13.07

On efficient parallelizing of the algorithms for discrete enumeration problems*

***E. V. Djukova*[1] *and A. G. Nikiforov*[2]**
[1]Dorodnicyn Computing Centre of the Russian Academy of Sciences, 40 Vavilova st., Moscow, Russia
[2]Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russia

Background: Approach to construction of efficient parallel algorithms for discrete enumeration problems is introduced in the previous works of the authors. This approach is based on statistical estimations for computational tasks size. The approach is demonstrated on dualization, which is an intractable problem and consists in enumeration of irreducible coverings of a given boolean matrix. The main disadvantage of formerly suggested parallel schemes for asympotically optimal dualization algorithms is time-costly tasks size estimation method which considers only the problem size.

Methods: A new parallel scheme has been developed for asymptotically optimal dualization algorithms, reducing time costs on statistical data collection. Statistical data are obtained via processing of a given matrix submatrices.

Results: Task distribution is performed according to schedule calculated in advance. For this purpose, a distribution of random variable, used for tasks size estimation, is fitted and the processor load level is optimized. A parallel scheme is applied to an asymptotically optimal algorithm RUNC-M.

Concluding Remarks: A new parallel scheme works not worse than the formerly suggested ones, demonstrates an almost maximal speedup and makes it possible to dualize matrices of big

*Работа частично поддержана грантами РФФИ № 13-01-00787-а и № 14-07-00819-а и грантом Президента РФ НШ-4908.2014.1.

size. However, this scheme is efficient only if the number of processors is significantly smaller than the number of matrix columns.

Keywords: *enumeration; dualization; irreducible covering of a boolean matrix; hypergraph transversal; asymptotically optimal algorithm; parallel computations; load balancing; strong scaling*

DOI: 10.21469/22233792.1.13.07

1 Введение

Одной из фундаментальных задач дискретной математики является дуализация. Ниже приведена ее матричная формулировка.

Дана булева матрица L размера $m \times n$. Набор H, состоящий из различных столбцов матрицы L, называется неприводимым покрытием, если он удовлетворяет двум условиям: (1) в подматрице L^H матрицы L, образованной столбцами набора H, не содержится строки вида $(0, 0, \ldots, 0)$; (2) подматрица L^H содержит каждую из строк вида $(1, 0, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, $\ldots$, $(0, 0, 0, \ldots, 1)$. Если набор столбцов H удовлетворяет условию (1), то он называется покрытием. Если набор столбцов H удовлетворяет условию (2), то он называется совместимым. Требуется построить множество $P(L)$ всех неприводимых покрытий матрицы L.

Дуализация имеет и другие эквивалентные формулировки. Приведем основные из них.

1. Дана конъюнктивная нормальная форма, реализующая монотонную булеву функцию F. Требуется построить сокращенную дизъюнктивную нормальную форму функции F.
2. Дан гиперграф G. Требуется перечислить все минимальные трансверсали гиперграфа G (двойственной задачей является задача перечисления всех минимальных вершинных покрытий гиперграфа G).

Дуализация возникает во многих областях дискретной математики (комбинаторике, теории гиперграфов, целочисленном программировании), в теории игр, в теории баз данных, в теории машинного обучения и т. д.

Асимптотические оценки типичных значений числа решений дуализации [1] показывают, что, как правило, это число растет экспоненциально с ростом размера входных данных. Поэтому дуализация относится к числу труднорешаемых перечислительных задач. Существует несколько подходов к оценке эффективности алгоритмов для перечислительных задач [1–3].

Говорят, что алгоритм работает с (квази)полиномиальной задержкой, если на каждом шаге строится ровно одно решение и сложность шага ограничивается (квази)полиномом от размера входных данных (для дуализации в матричной формулировке — это размер матрицы L). Алгоритмы дуализации с (квази)полиномиальной задержкой удалось построить только для некоторых частных случаев (например, когда в каждой строке исходной матрицы не более двух единиц [3]). Таким образом, статус дуализации в плане полиномиальной разрешимости неизвестен.

В [1, 4] предложен подход к построению асимптотически оптимальных алгоритмов дуализации. В дальнейшем этот подход получил развитие в [5–8]. На каждом шаге асимптотически оптимального алгоритма строится набор столбцов матрицы, удовлетворяющий условию совместимости (2). В отличие от алгоритма дуализации с полиномиальной задержкой асимптотически оптимальный алгоритм может делать полиномиальные «лиш-

ние» шаги, причем доля таких шагов стремится к нулю для почти всех матриц L размера $m\times n$ при $m, n \to \infty$. На «лишнем» шаге либо строится набор столбцов матрицы, не являющийся покрытием, либо строится набор столбцов, найденный ранее. Проверка на повторяемость построенного набора столбцов осуществляется за полиномиальное время от размеров матрицы.

Асимптотически оптимальные алгоритмы являются лидерами по скорости счета среди других известных алгоритмов дуализации. В [9, 10] показано, что наиболее быстро среди асимптотически оптимальных алгоритмов работают алгоритмы RUNC-M и PUNC [10], не делающие «повторных» шагов.

В силу того что число решений дуализации, как правило, растет экспоненциально с ростом размеров входных данных, актуальным является использование параллельных вычислений. Существуют простые и очевидные схемы распараллеливания асимптотически оптимальных алгоритмов дуализации, основным недостатком которых является неравномерная загрузка процессоров, что приводит и к недостаточному ускорению времени работы параллельного алгоритма по сравнению с его последовательной версией. Схема распараллеливания определяется способом выбора вычислительных подзадач и способом распределения этих подзадач между процессорами.

Следует отметить, что за рубежом при создании параллельных алгоритмов дуализации первостепенное внимание уделяется теоретическим оценкам их сложности в зависимости от числа используемых процессоров [11, 12], причем, как правило, строятся алгоритмы, ориентированные на частные случаи, например, когда число единиц в каждой строке исходной матрицы ограничено некоторой небольшой константой.

В [13] предложен подход к построению эффективных в практическом плане параллельных асимптотически оптимальных алгоритмов дуализации. Опишем разработанную в [13] B-схему распараллеливания.

Пусть H — неприводимое покрытие матрицы L, состоящее из столбцов с номерами $j_1, \ldots, j_r$, где $j_1 < \cdots < j_r$. Тогда H назовем j_1-неприводимым покрытием. Подзадача с номером j, $j \in \{1, \ldots, n\}$, состоит в построении множества $P_j(L)$ всех j-неприводимых покрытий матрицы L. Объемы подзадач определяются величинами $\nu_j(L) = |P_j(L)|/|P(L)|$, $j \in \{1, \ldots, n\}$.

B-схема имеет статический характер: распределение подзадач происходит по заранее составленному «расписанию». Статистическая обработка экспериментов показывает, что случайная величина, использующаяся для оценки $\nu_j(L), j \in \{1, \ldots, n\}$, подчиняется бета-биномиальному закону, параметры которого вычисляются при помощи метода максимального правдоподобия по выборке из случайных матриц размера $m\times n$. Для составления «расписания» решается задача оптимизации уровня загрузки процессоров.

B-схема демонстрирует, как правило, достаточно равномерную загрузку процессоров и высокое ускорение времени работы при увеличении числа процессоров. Однако вычисление оценок для $\nu_j(L)$ в этой схеме требует многократного решения задачи дуализации для случайных матриц, имеющих одинаковый размер с матрицей L, и эти оценки недостаточно точны.

Основным результатом данной работы является разработка S-схемы распараллеливания асимптотически оптимальных алгоритмов дуализации. Эта схема отличается от B-схемы методом получения оценок для $\nu_j(L)$, $j \in \{1, \ldots, n\}$, который является менее трудоемким и учитывает не только размеры матрицы. Метод основан на обработке случайных подматриц данной матрицы, подматрицы имеют размер $r\times n$, где r является параметром и не превосходит m. Выявлено, что при параметре r, равном $m/2$, полученные оценки

являются достаточно точными с точки зрения критерия Хи-квадрат. Работа S-схемы продемонстрирована на примере алгоритма RUNC-M [10], который является модификацией асимптотически оптимального алгоритма ОПТ [8].

При тестировании S-схемы исследуется ее сильная масштабируемость (зависимость основных показателей работы параллельного алгоритма от числа процессоров при фиксированном размере задачи). Показано, что S-схема демонстрирует такие же показатели сильной масштабируемости, как B-схема, и в то же время позволяет обрабатывать матрицы еще бóльших размеров за счет более быстрого вычисления оценок для объемов подзадач.

2 Описание асимптотически оптимального алгоритма дуализации RUNC-M

В данном разделе приводится описание асимптотически оптимального алгоритма дуализации RUNC-M [10]. Подробно описана структура дерева решений, которое строит данный алгоритм.

Обозначим через M_{mn} множество булевых матриц размера $m\times n$, а через J_u множество $\{1, 2, \ldots, u\}$. Пусть $L = (a_{ij}) \in M_{mn}$. В данном разделе будем отождествлять набор столбцов (строк) матрицы L с набором их номеров.

Будем говорить, что столбец j (столбец с номером j) покрывает строку i (строку с номером i) матрицы L, если $a_{ij} = 1$.

Строка i матрицы L является опорной для пары (H, j), $j \in H$, если $a_{ij} = 1$ и $a_{ij} = 0\ \forall u \in H \setminus \{j\}$. Множество всех опорных строк для (H, j) обозначим через $S(H, j)$. Очевидно, набор H является совместимым тогда и только тогда, когда $S(H, j) \neq \varnothing\ \forall j \in H$.

Говорят, что столбец j матрицы L совместим с совместимым набором H, если набор $H \cup \{j\}$ совместимый. Очевидно, столбец j не совместим с совместимым набором H тогда и только тогда, когда $\exists u \in H$ такой, что столбец j покрывает все строки из $S(H, u)$.

Асимптотически оптимальный алгоритм дуализации RUNC-M перечисляет с полиномиальной задержкой $O(qmn), q = \min\{m, n\}$, некоторое подмножество совместимых наборов столбцов матрицы L, содержащее множество $P(L)$. Алгоритм строит дерево решений, совершая его обход в глубину. Построение одной висячей вершины — это шаг алгоритма.

Вершина (H, R, C) дерева решений описывается совместимым набором столбцов H, набором строк R и набором столбцов C. В висячей вершине имеет место один из двух случаев: (1) $R = \varnothing$; (2) $R \neq \varnothing, C = \varnothing$. В первом случае H — неприводимое покрытие. Во втором случае висячая вершина соответствует «лишнему» шагу. Корню дерева соответствует $(H = \varnothing, R = J_m, C = J_n)$. Пусть построена внутренняя (не висячая) вершина (H, R, C), тогда переход к следующей построенной вершине будет осуществляться путем добавления к H столбца из C и удаления некоторых строк и столбцов из R и C соответственно.

Пусть построена внутренняя вершина (H_0, R_0, C_0). Тогда при построении следующей вершины к H_0 добавляется первый по порядку столбец из $C_0^{\min} = \{j \in C_0 | a_{ij} = 1\}$, где $i \in R_0$ — номер строки с наименьшей суммой $\sum_{j\in C_0} a_{ij}$ (если таких строк несколько, то выбирается строка с наименьшим номером среди них).

Для того чтобы схемы распараллеливания, описанные далее, были применимы к алгоритму RUNC-M, требуется его немного модифицировать: на первом ярусе дерева решений, или когда глубина рекурсии равна нулю, вместо множества $C_0^{\min}$ берется множество $C_0 = J_n$.

Алгоритм 1 BuildSubtreeRUNCM

Вход: L, H_0, R_0, C_0;
Выход: $\varnothing$;
1: $C_0^{\min} = \{j \in C_0 | a_{ij} = 1\}$, где $i \in R_0$ — номер строки с наименьшей суммой $\sum_{j \in C_0} a_{ij}$
2: **для всех** $j \in C_0^{\min}$
3: $R \leftarrow R_0$
4: $C_0 \leftarrow C_0 \setminus \{j\}$
5: $C \leftarrow C_0$
6: $H \leftarrow H_0 \cup \{j\}$
7: Удалить из R строки, покрытые столбцом j
8: **если** $R = \varnothing$ **то**
9: Сохранить набор H, который является неприводимым покрытием
10: **иначе**
11: Удалить из C не совместимые с набором H столбцы
12: BuildSubtreeRUNCM(L, H, R, C)

Опишем рекурсивную процедуру BuildSubtreeRUNCM(L, H, R, C) (см. Алгоритм 1) построения поддерева решений. Для запуска алгоритма эту функцию следует вызывать с параметрами $H = \varnothing, R = J_m, C = J_n$. Отметим, что все аргументы процедуры передаются по значению или копируются.

Настоящая реализация алгоритма RUNC-M написана на языке `C++` с интенсивным использованием побитовых операций. Кроме того, для некоторых частей алгоритма используется динамический выбор функций для минимизации числа операций (например, этот прием используется при удалении несовместимых строк).

3 Оценки для объемов подзадач, используемые в S-схеме

В данном разделе описаны способы оценки объемов подзадач или величин $\nu_j(L) = |P_j(L)|/|P(L)|$, $j \in J_n$, используемые соответственно в B-схеме [13] и S-схеме.

В B-схеме на пространстве равновероятных элементарных событий $\Omega = \{(L, H) \,|\, L \in M_{mn}, H \in P(L)\}$ вводится случайная величина $\eta(L, H)$, равная j, если $H \in P_j(L), j \in J_n$. При помощи критерия Хи-квадрат проверяется гипотеза о виде распределения $H_0 : f(j) = \psi_{\alpha\beta}(j)$, где $f(j)$ — вероятность события $\eta(L, H) = j$, а $\psi_{\alpha\beta}(j)$ — функция вероятности бета-биномиального распределения с параметрами α и β, которые оцениваются при помощи метода максимального правдоподобия. B-схема использует величины $\psi_{\alpha\beta}(j)$ в качестве приближенного значения искомой величины $\nu_j(L), j \in J_n$. Эта схема обладает двумя основными недостатками: оценка для $\nu_j(L)$ одна и та же для всех матриц данного размера, и для ее вычисления требуется многократно решить задачу дуализации для случайных матриц из M_{mn}.

Теперь опишем S-схему.

Пусть $L \in M_{mn}$ и $r \leqslant m$. Через W_m^r обозначим множество всех подмножеств мощности r множества J_m. Пусть $w \in W_m^r$, тогда через L^w обозначим подматрицу матрицы L, составленную из строк матрицы L с номерами из w.

Пусть $\Omega_r = \{(L^w, H) \,|\, w \in W_m^r, H \in P(L)\}$ — пространство равновероятных элементарных событий. На указанном пространстве определим случайную величину $\eta_r(L^w, H)$,

Таблица 1 Значения пар $(Z_r(\boldsymbol{x}), \gamma_r^*(\boldsymbol{x}))$ для критерия Хи-квадрат

$r \setminus m{\times}n$	30×120	40×120	50×100	70×70
10	$(159, < 10^{-4})$	$(167, < 10^{-4})$	$(235, < 10^{-4})$	$(382, < 10^{-4})$
13	$(99, < 10^{-4})$	$(132, < 10^{-4})$	$(157, < 10^{-4})$	$(234, < 10^{-4})$
15	$(77, 0{,}0134)$	$(112, < 10^{-4})$	$(117, < 10^{-4})$	$(187, < 10^{-4})$
18	$(74, 0{,}028)$	$(90, 0{,}0002)$	$(96, < 10^{-4})$	$(147, < 10^{-4})$
20	$(60, 0{,}0815)$	$(63, 0{,}0546)$	$(89, < 10^{-4})$	$(131, < 10^{-4})$
25	$(54, 0{,}315)$	$(60, 0{,}0876)$	$(50, 0{,}1382)$	$(85, < 10^{-4})$
30	—	—	—	$(68, 0{,}0001)$
35	—	—	—	$(54, 0{,}0478)$

которая равна $j, j \in J_n$, если $H \in P_j(L)$. Через $f_r(j)$ обозначим вероятность события $\eta_r(L^w, H) = j$.

Предлагается использовать величину $f_r(j)$ в качестве приближенного значения искомой величины $\nu_j(L), j \in J_n$. Встает вопрос, при каких r указанные оценки являются достаточно точными. С одной стороны, число r должно быть как можно меньшим, чтобы время получения оценок было относительно невелико. С другой стороны, оценки должны быть достоверными.

Пусть $\boldsymbol{x} = (x_1, \ldots, x_N)$ — выборка из распределения $f_r(j)$. Для проверки статистической гипотезы $H_0 : f_r(j) = \nu_j(L)$ о виде распределения случайной величины $\eta_r(L^w, H)$ предлагается использовать критерий Хи-квадрат со статистикой

$$Z_r(\boldsymbol{x}) = N \sum_{j=1}^{n} \frac{(f^*(j) - \nu_j(L))^2}{\nu_j(L)},$$

где $f^*(j)$ — доля элементов выборки $\boldsymbol{x} = (x_1, \ldots, x_N)$, равных j.

Достигнутым уровнем значимости критерия Хи-квадрат называется величина $\gamma_r^*(\boldsymbol{x}) = 1 - \chi^2_{n-1}(Z_r(\boldsymbol{x}))$, где χ^2_{n-1} — функция распределения Хи-квадрат с $(n-1)$ степенями свободы. Близость значения $\gamma_r^*(\boldsymbol{x})$ к 0 говорит о том, что гипотезу H_0 вероятнее всего следует отклонить.

Для получения выборки $\boldsymbol{x} = (x_1, \ldots, x_N)$ из распределения $f_r(j)$ построим t случайных подматриц L^w матрицы L размера $r{\times}n$. Выберем N пар $(L^w, H), H \in P(L^w)$, и из значений случайной величины $\eta_r(L^w, H)$ сформируем выборку.

Проведем эксперимент. Для каждой из конфигураций 30×150, 40×120, 50×100 и 70×70 выберем по 20 случайных матриц. Для каждой матрицы сформируем выборку $\boldsymbol{x} = (x_1, \ldots, x_N)$ из распределения $f_r(j)$, где $N = 1000$. В табл. 1 приведены медианные значений статистики $Z_r(\boldsymbol{x})$ и достигаемых уровней значимости $\gamma_r^*(\boldsymbol{x})$. На рис. 1 приведены графики величин $\nu_j(L)$ и $f_r^*(j)$.

Согласно табл. 1 минимальное значение r, при котором достигнутый уровень значимости $\gamma_r^*(\boldsymbol{x})$ не является пренебрежимо малым, равняется $m/2$. На примере конфигурации 30×150 можно заметить, что при $r = 15$ имеет место «фазовый переход»: при пересечении этой точки функция $Z_r(\boldsymbol{x})$ начинает стабилизироваться. Это говорит о том, что дальнейшее увеличение r не принесет существенного выигрыша в приближении $\nu_j(L)$.

4 Распределение вычислительных заданий между процессорами

Пусть $L \in M_{mn}$ и пусть дано $p \leqslant n$ процессоров. Пусть j-я подзадача обрабатывается процессором с номером N_j. Вектор $\boldsymbol{N^p} = (N_1, \ldots, N_n)$ назовем расписанием. Уровнем

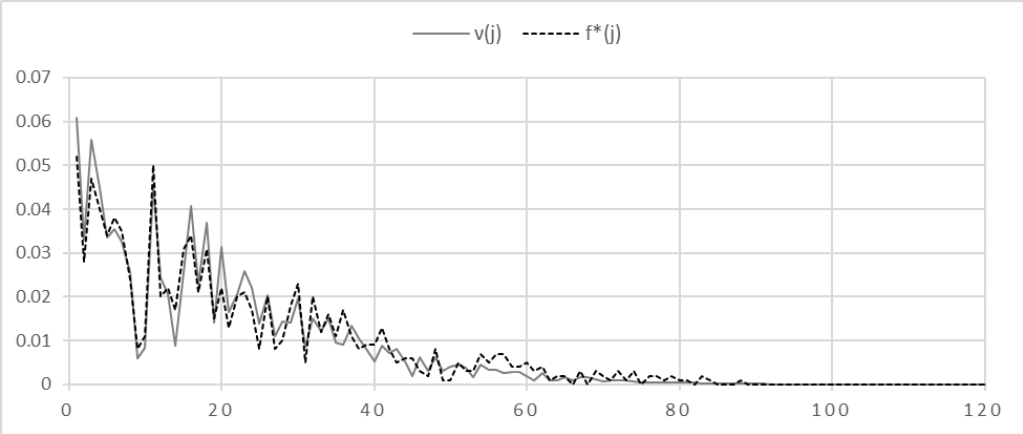


Рис. 1 Графики $\nu_j(L)$ и $f_r^*(r)$ как величин, зависящих от j, при $m = 30$, $n = 120$ и $r = 15$

загрузки k-го процессора назовем величину

$$\sigma_k(\boldsymbol{N^p}) = \sum_{j \in J_n : N_j = k} \nu_j(L).$$

Для эффективного распределения вычислительных заданий между процессорами, требуется решить задачу минимизации уровня загрузки процессоров

$$\sigma(\boldsymbol{N^p}) = \max_{k \in J_p} \sigma_k(\boldsymbol{N^p}) \to \min_{\boldsymbol{N^p}}. \tag{1}$$

Ниже приведено описание процедуры 2, которая ищет приближенное решение задачи 1 при помощи жадного алгоритма. На вход этой процедуры подается число процессоров p, число столбцов n матрицы L и вектор $\widetilde{\boldsymbol{\nu}} = (\widetilde{\nu}_1, \ldots, \widetilde{\nu}_n)$, состоящий из оценок для величин $\nu_j(L)$. Способы получения оценок $\widetilde{\nu}_j$ описаны ранее в этой работе.

Алгоритм 2 DistributeTasks

Вход: $p, n, \widetilde{\boldsymbol{\nu}}$;
Выход: $\boldsymbol{N^p}$;

1: **для всех** $k \in \{1, \ldots, p\}$
2: $\quad \sigma_k \leftarrow 0$
3: **для** $j \in \{1, \ldots, n\}$
4: $\quad k_0 \leftarrow \arg\min_{k \in J_p} \sigma_k$
5: $\quad N_j \leftarrow k_0$
6: $\quad \sigma_k \leftarrow \sigma_k + \widetilde{\nu}_j$

5 Тестирование S-схемы распараллеливания

В данном разделе даны описания среды тестирования и исследуемых показателей работы параллельных алгоритмов, приведены результаты сравнения S-схемы и B-схемы и результаты дополнительного тестирования S-схемы на больших матрицах.

Таблица 2 Сравнение схем распараллеливания при $m = 65$ и $n = 80$

p	B-схема			S-схема		
	$T(p)$	$\sigma(p)$	$s(p)$	$T(p)$	$\sigma(p)$	$s(p)$
1	17,40	1,000	1,000	18,35	1,000	1,000
2	9,01	0,500	0,514	9,40	0,500	0,502
4	5,30	0,250	0,291	4,92	0,250	0,261
8	2,71	0,125	0,147	2,52	0,125	0,135
16	1,55	0,079	0,090	1,62	0,084	0,087
32	1,55	0,079	0,090	1,61	0,089	0,087
64	1,55	0,079	0,090	1,61	0,089	0,087

Таблица 3 Сравнение схем распараллеливания при $m = 80$ и $n = 65$

p	B-схема			S-схема		
	$T(p)$	$\sigma(p)$	$s(p)$	$T(p)$	$\sigma(p)$	$s(p)$
1	26,1	1,000	1,000	26,2	1,000	1,000
2	13,7	0,500	0,514	13,8	0,500	0,507
4	7,17	0,250	0,271	7,01	0,250	0,255
8	3,87	0,125	0,140	3,79	0,126	0,137
16	2,83	0,102	0,114	3,13	0,086	0,123
32	2,83	0,102	0,114	3,13	0,092	0,123
64	2,83	0,102	0,114	3,13	0,092	0,123

Тестирование проводилось на суперкомпьютере IBM Blue Gene/P, располагающегося в МГУ им. М. В. Ломоносова в здании факультета Вычислительной математики и кибернетики и являющегося массивно-параллельной вычислительной системой. Каждый вычислительный узел включает в себя четырехъядерный процессор PowerPC 450 (850 МГц), 2 ГБ общей памяти и сетевые интерфейсы. При запуске вычислительных заданий использовался режим виртуальных вычислительных узлов (VN (virtual network) режим). В этом режиме на каждом вычислительном узле запущено четыре MPI (message passing interface) процесса, которые делят между собой доступные ресурсы.

Через p обозначим число процессоров, через $T_k(p)$ — время (в секундах) работы k-го процессора параллельного алгоритма при использовании p процессоров. Пусть $T(p) = \max\limits_k T_k(p)$ и $T_\Sigma(p) = \sum_k T_k(p)$. Достигнутым уровнем загрузки k-го процессора назовем величину $s_k(p) = T_k(p)/T_\Sigma(p)$. Исследуются три показателя:

(1) ускорение алгоритма $S(p) = T(1)/T(p)$;
(2) равномерность загрузки процессоров $E(p) = S(p)/p$;
(3) достигнутый уровень загрузки $s(p) = \max\limits_k s_k(p)$.

Ускорение, соответствующее линейной функции $S(p) = p$ при $p \geqslant 1$, является практически максимальным. Близость функции $E(p)$ к единице свидетельствует о равномерной загрузке процессоров. Показатель $s(p)$ является аналогом показателя уровня загрузки процессоров $\sigma(\boldsymbol{N^p})$, определенного ранее в этой работе.

Поскольку используемые в B-схеме оценки для $\nu_j(L)$ одинаковы для всех матриц данного размера, далее эти оценки считаются известными во время работы параллельного

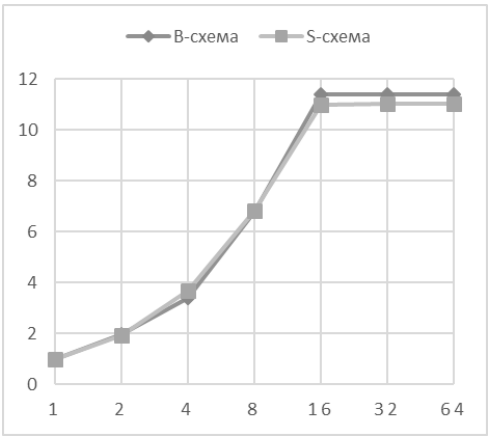
Таблица 4 Сравнение схем распараллеливания при $m = 80$ и $n = 80$

p	B-схема			S-схема		
	$T(p)$	$\sigma(p)$	$s(p)$	$T(p)$	$\sigma(p)$	$s(p)$
1	34,4	1,000	1,000	36,5	1,000	1,000
2	17,3	0,500	0,502	18,8	0,500	0,508
4	10,1	0,250	0,286	9,94	0,250	0,257
8	5,10	0,125	0,147	5,35	0,125	0,137
16	3,03	0,078	0,091	3,33	0,074	0,085
32	3,03	0,078	0,091	3,32	0,076	0,085
64	3,03	0,078	0,091	3,32	0,076	0,085

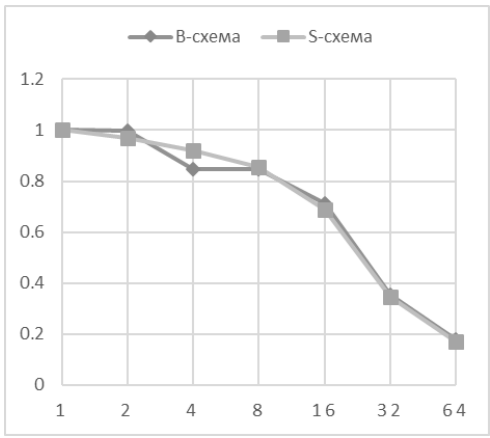
алгоритма. В S-схеме, напротив, оценки для объемов подзадач подсчитываются во время работы параллельного алгоритма.

Сравнение B-схемы и S-схемы проводится на матрицах размера 65×80, 80×65 и 80×80. Матрицы бóльших конфигураций при сравнении не рассматриваются ввиду трудоемкости получения оценок для B-схемы. Результаты представлены в табл. 2–4. На рис. 2 представлены графики функций $S(p)$ и $E(p)$ при $m = n = 80$.

Из указанных таблиц и графиков следует, что обе схемы демонстрируют практически одинаковое ускорение $S(p)$ и равномерность загрузки $E(p)$. При этом достигнутый уровень загрузки $s(p)$ у S-схемы наиболее низкий и $s(p) \approx \sigma(p)$, что свидетельствует о качественной балансировке нагрузки.



(*а*) График $S(p)$ при $m = 80$ и $n = 80$



(*б*) График $E(p)$ при $m = 80$ и $n = 80$

Рис. 2 Сравнение схем распараллеливания

Для каждого из размеров матриц можно указать число p^* такое, что при $p \leqslant p^*$ обе схемы эффективны: $S(p) \approx p$ и $E(p) \approx 1$. Например, при $m = n = 80$ число p^* равно 16. При $p > p^*$ значения $T(p)$ «стабилизируются». Это связано с тем, что распараллеливание происходит на первом ярусе дерева решений, которое строит алгоритм RUNC-M. При таком подходе объемы подзадач сильно различаются, поэтому их принципиально невозможно равномерно распределить между большим числом процессоров.

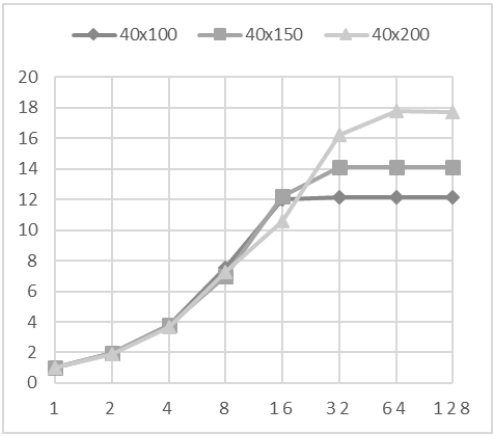
Вычисление оценок для объемов подзадач, используемых в S-схеме, гораздо менее трудоемкое, чем в B-схеме. Поэтому S-схема применима и для больших матриц. Это демон-

Таблица 5 Время работы $T(p)$ для S-схемы

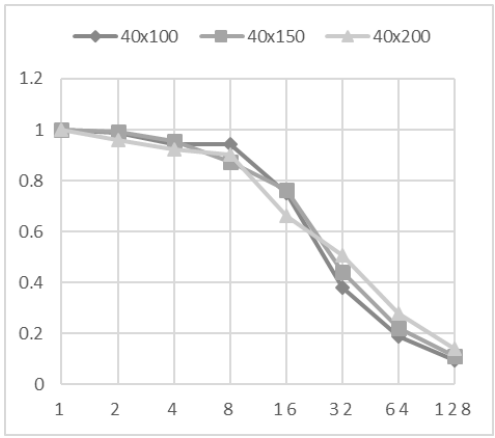
$m\times n \setminus p$	1	2	4	8	16	32	64	128
30×100	3,95	2,03	1,05	0,59	0,37	0,32	0,32	0,32
30×150	39,1	20,0	10,4	5,21	3,46	2,32	2,33	2,32
30×200	231	116	61,5	32,2	18,8	13,8	13,8	13,8
40×100	11,5	5,83	3,05	1,53	0,96	0,95	0,95	0,95
40×150	133	67,1	34,8	19,1	10,9	9,44	9,43	9,43
40×200	654	328	177	90,5	61,8	40,4	36,8	36,8

стрируется на матрицах размера $m\times n$, где $m \in \{30, 40\}$ и $n \in \{100, 150, 200\}$. Значение параметра r полагалось равным $m/2$.

В табл. 5 приведено время работы $T(p)$ параллельного алгоритма дуализации, основанного на S-схеме, при различных m, n и p. На рис. 3 приведены графики $S(p)$ и $E(p)$. На рис. 4 приведена столбчатая диаграмма для достигнутых уровней загрузки $s_k(p), k \in J_p$ при $p = 16$ и 32.



(*а*) График $S(p)$ при $m = 40$



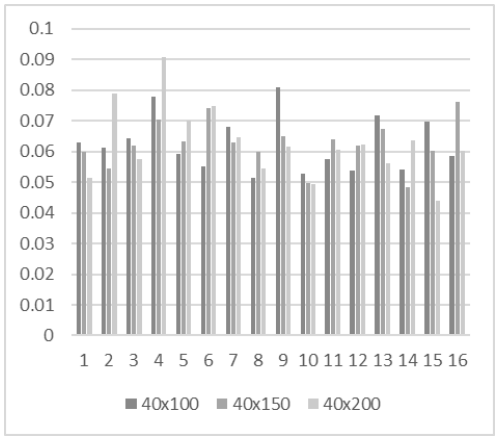
(*б*) График $E(p)$ при $m = 40$

Рис. 3 Сильная масштабируемость S-схемы

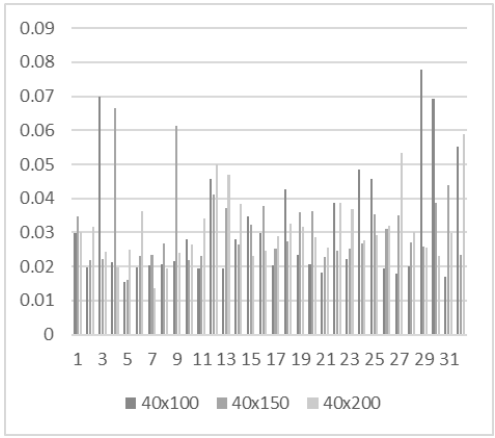
S-схема демонстрирует практически линейное ускорение и высокий уровень загрузки при $p \leqslant p^*$. Например, $p^* = 32$ при $m = 40$ и $n = 200$. Согласно рис. 4 для матрицы 40×200 достигнутый уровень загрузки при $p = 32$ для некоторых процессоров значительно превышает среднее значение этого показателя, что может быть результатом недостаточного качества оценки $f_r^*(j)$ или неоптимальности построенного жадным алгоритмом расписания.

6 Заключение

В данной работе развит предложенный в [13] подход к построению параллельных алгоритмов для дискретных перечислительных задач. Подход основан на статистических оценках объемов вычислительных подзадач. Распределение вычислительных подзадач осуществляется согласно заранее составленному расписанию. Для составления указанного



(*а*) График $s_k(16)$ при $m = 40$



(*б*) График $s_k(32)$ при $m = 40$

Рис. 4 Достигнутый уровень загрузки в S-схеме

расписания определяется вид распределения случайной величины, использующейся для оценки объемов подзадач, и оптимизируется уровень загрузки процессоров. В рамках рассматриваемого подхода разработана новая менее трудоемкая схема распараллеливания асимптотически оптимальных алгоритмов дуализации.

Работа схемы продемонстрирована на примере распараллеливания алгоритма RUNC-M [10], который в настоящее время является лидером по скорости счета среди алгоритмов дуализации. Предлагаемый подход к построению параллельных алгоритмов дуализации обеспечивает высокую точность оценок для объемов подзадач, что при определенных условиях приводит и к высоким показателям эффективности параллельного алгоритма. Однако рассматриваемый подход не эффективен при большом числе процессоров, поскольку вычислительные подзадачи имеют существенно разные размеры (распараллеливание происходит на первом ярусе дерева решений, которое строит асимптотически оптимальный алгоритм дуализации).

Литература

[1] *Дюкова Е. В.* Об асимптотически оптимальном алгоритме построения тупиковых тестов // Докл. АН СССР, 1977. Т. 223. № 4. С. 527–530.

[2] *Johnson D. S., Yannakis M., Papadimitriou C. H.* On generating all maximal independent sets // Inform. Processing Lett., 1988. Vol. 27. P. 119–123.

[3] *Khachiyan L., Fredman M.* On the complexity of dualization of monotone disjunctive normal forms // J. Algorithms, 1996. Vol. 21, No. 3. P. 618–628.

[4] *Дюкова Е. В.* Асимптотически оптимальные тестовые алгоритмы в задачах распознавания // Пробл. кибернетики, 1982. Т. 1. № 39. С. 165–199.

[5] *Djukova E. V., Zhuravlev Y. I.* Discrete methods of information analysis in recognition and algorithm synthesis // Pattern Recogn. Image Anal., 1997. Vol. 7. No. 2. P. 192–207.

[6] *Дюкова Е. В., Журавлев Ю. И.* Дискретный анализ признаковых описаний в задачах распознавания большой размерности // Ж. вычисл. матем. и матем. физ., 2000. Т. 40. № 8. С. 1264–1278.

[7] *Дюкова Е. В.* О сложности реализации дискретных (логических) процедур распознавания // Ж. вычисл. матем. и матем. физ., 2004. Т. 44. № 3. С. 551–561.

[8] *Дюкова Е. В., Инякин А. С.* Асимптотически оптимальное построение тупиковых покрытий целочисленной матрицы // Математические вопросы кибернетики, 2008. Т. 17. С. 235–246.

[9] *Murakami K., Uno T.* Efficient algorithms for dualizing large-scale hypergraphs // Discrete Appl. Math., 2014. Vol. 170. P. 83–94.

[10] *Дюкова Е. В., Прокофьев П. А.* Построение и исследование новых асимптотически оптимальных алгоритмов дуализации // Машинное обучение и анализ данных, 2014. Т. 1. № 8. С. 1048–1067.

[11] *Khachiyan L., Boros E., Elbassioni K., Gurvich V.* A new algorithm for the hypergraph transversal proble // Computing and combinatorics / Ed. L. Wang. — Lecture notes in computer science ser. — Springer, 2005. Vol. 3595. P. 767–776.

[12] *Khachiyan L., Boros E., Gurvich V., Elbassioni K.* Computing many maximal independent sets for hypergraphs in parallel // Parallel Processing Lett., 2007. Vol. 17, No. 2. P. 141–152.

[13] *Дюкова Е. В., Никифоров А. Г., Прокофьев П. А.* Статистически эффективная схема распараллеливания алгоритмов дуализации // Машинное обучение и анализ данных, 2014. Т. 1. № 7. С. 846–853.

Поступила в редакцию 16.06.2015

References

[1] Djukova, E. V. 1977. On an asympotically optimal algorithm for constructing irredundant tests. *Dokl. Akad. Nauk SSSR* 223(4):527–530.

[2] Johnson, D. S., M. Yannakis, and C. H. Papadimitriou. 1988. On generating all maximal independent sets. *Inform. Processing Lett.* 27:119–123.

[3] Khachiyan L., and M. Fredman. 1996. On the complexity of dualization of monotone disjunctive normal forms. *J. Algorithms* 21(3):618–628.

[4] Djukova, E. V. 1982. Asimptoticheski optimal'nye testovye algoritmy v zadachakh raspoznavaniya. *Problemy Kibernetiki* 1(39):165–199.

[5] Djukova, E. V., and Y. I. Zhuravlev. 1997. Discrete methods of information analysis in recognition and algorithm synthesis. *Pattern Recogn. Image Anal.* 7(2):192–207.

[6] Djukova, E. V., and Y. I. Zhuravlev. 2000. Diskretnyy analiz priznakovykh opisaniy v zadachakh raspoznavaniya bol'shoy razmernosti. *J. Vychisl. Matem. i Matem. Fiz.* 40(8):1264–1278.

[7] Djukova, E. V. O slozhnosti realizatsii diskretnykh (logicheskikh) protsedur raspoznavaniya. *J. Vychisl. Matem. i Matem. Fiz.* 44(3):551–561.

[8] Djukova, E. V., and A. S. Inyakin. 2008. Asymptoticheski optimal'noe postroenie tupikovykh pokrytiy tselochislennoy matritsy. *Matematicheskie Vorposy Kibernetiki* 17:235–246.

[9] Murakami, K., and T. Uno. 2014. Efficient algorithms for dualizing large-scale hypergraphs. *Discrete Appl. Math.* 170:83–94.

[10] Djukova, E. V., and P. A. Prokofyev. 2014. Construction and investigation of new asymptotically optimal algorithms for dualization. *Machine Learning Data Anal.* 1(8):1048–1067.

[11] Khachiyan, L., E. Boros, K. Elbassioni, and V. Gurvich. 2005. A new algorithm for the hypergraph transversal proble. *Computing and combinatorics*. Ed. L. Wang. Lecture notes in computer science ser. Springer. 3595:767–776.

[12] Khachiyan, L., E. Boros, V. Gurvich, and K. Elbassioni. 2007. Computing many maximal independent sets for hypergraphs in parallel. *Parallel Processing Lett.* 17(2):141–152.

[13] Djukova, E. V., A. G. Nikiforov, and P. A. Prokofyev. 2014. Statistically efficient parallel scheme for dualization algorithms. *Machine Learning Data Anal.* 1(7):846–853.

Received June 16, 2015

Параллельные тексты в задаче дешифровки древнерусских знаменных песнопений*

И. В. Бахмутова, В. Д. Гусев, Л. А. Мирошниченко, Т. Н. Титкова
gusev@math.nsc.ru, luba@math.nsc.ru
Новосибирск, Институт математики им. С. Л. Соболева СО РАН

Предложен новый компьютерно-ориентированный подход к проблеме нотолинейной реконструкции (дешифровки) древнерусских церковных песнопений XVI–XVII вв., представленных в знаменной форме записи. В основе подхода лежит анализ двознаменников — певческих книг, записанных в виде трех параллельных (синхронизованных между собой) текстов: знаменного, нотолинейного и стихотворного (старославянского). Введены понятия инвариантов и квазиинвариантов знаменного распева. Разработан алгоритм выделения их из обучающего материала (двознаменники) и использования для целей дешифровки. Получены оценки эффективности подхода на независимом контрольном материале. Основным достоинствам подхода является ориентация его на общий случай беспометной нотации.

Ключевые слова: *параллельные тексты; знаменные песнопения; двознаменники; дешифровка; инварианты; квазиинварианты*

DOI: 10.21469/22233792.1.13.08

Parallel texts in the problem of deciphering of ancient Russian chant*

I. V. Bakhmutova, V. D. Gusev, L. A. Miroshnichenko, and T. N. Titkova
Sobolev Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences, 4 Acad. Koptyug avenue, Novosibirsk, Russia

The ancient Russian chants of XII–XVII centuries are presented in the neume writing form. The problem of chant translation into modern note writing is of deciphering character and, in the general case (chants without special marks that explain their singing value), is not yet solved. The number of "unreadable" ancient hymnals runs into the hundreds. The main difficulties of deciphering are connected with the polysemy of correspondence "neume–note." The known examples of deciphering are few in number, made by manually, and refer to the separate hymn. The authors develop a new computer-oriented approach to the solution of this problem using the dvoyeznamenniks of the end of XVII – beginning of XVIII centuries where the chants are written in three (synchronized between each other) parallel text: in neumes, in notes, and in old Slavonic verses. The emphasis places on revealing in texts of dvoyeznamenniks not very long repeating chains of neumes that are interpreted either unambiguously (invariants) or with admissible deviations (quasi-invariants). On the basis of rather extensive learning material, the electronic dictionaries of invariants and quasi-invariants were constructed. The algorithm of deciphering of neumatic notation using these dictionaries was developed. The experiments on the control material have shown that at this stage (without appellation to the structural organization of neumatic hymnals), these dictionaries provide the deciphering of 60%–70% of neumatic text. The main features of the presented approach are: use of dvoyeznamenniks of

*Работа выполнена при финансовой поддержке РФФИ, проект № 13-07-00400.

the golden age period of Russian chant in different genres and orientation, in general, toward the neumatic notation without special marks from XVI–XVII centuries.

Keywords: *parallel texts; the ancient Russian chants; dvoyeznamenniks; deciphering; invariants; quasi-invariants*

DOI: 10.21469/22233792.1.13.08

1 Введение

Древнерусские церковные песнопения XII–XVIII вв. представлены преимущественно в знаменной форме записи. Проблема перевода их в современную нотолинейную форму носит дешифровочный характер и в общем случае (песнопения раннего и среднего периодов) остается нерешенной. Знамена (или крюки) — специальные знаки, служащие для передачи мелодии. Они интерпретируются цепочками нот разной длины (обычно от одного до пяти нотных знаков). Общие сведения о знаменном (или крюковом) пении можно найти в [1]. По оценкам, приведенным в [2], количество известных певческих рукописей измеряется на данный момент более чем тысячью экземпляров, из них читаемыми (с существенными оговорками [3, 4]) считаются лишь около половины.

Особый интерес в плане дешифровки представляют так называемые *двознаменники* (см. [1, гл. 12]). Это певческие книги конца XVII – начала XVIII вв., содержащие песнопения, записанные в виде трех синхронизованных *параллельных текстов*: знаменного, нотолинейного и стихотворного (старославянский язык). Количество известных двознаменников невелико (порядка десятка [1]). Именно они лежат в основе предлагаемого авторами подхода к дешифровке. Часть из них используется для обучения, другая — для контроля. Этому предшествовала достаточно объемная и кропотливая работа по переводу двознаменников в цифровую форму, поскольку программ распознавания знаменной нотации на данный момент, насколько известно авторам, не существует.

Немногочисленные известные примеры дешифровки знаменной нотации сделаны вручную и касаются отдельных песнопений (или узких классов песнопений), эволюцию которых можно проследить по архивным материалам в течение достаточно длительного периода. Существенную роль при этом играет наличие графически близких византийских версий. Подготовленные авторами электронные версии двознаменников охватывают песнопения разных жанров и позволяют использовать для дешифровки новые подходы, превышающие по своей трудоемкости возможности исследователей, работающих вручную. Эти подходы основаны на введенных авторами понятиях *инвариантов и квазиинвариантов* знаменного распева.

Целью работы является описание алгоритма дешифровки, использующего параллельные тексты двознаменников, и оценка его эффективности. Проводится анализ ошибок и рассматриваются возможности дальнейшего продвижения.

2 Краткие сведения о знаменном распеве

В истоках знаменного распева лежит древнегреческое церковное пение на 8 ладов (гласов): дорийский, фригийский и т. д. (так называемая *система осмогласия* [1]). Начало системе положил обычай в каждый из восьми дней Пасхи исполнять песнопения на особый напев. Восьмидневный цикл напевов, которые хор исполнял в унисон, был распространен затем на 8 недель, а одноголосный (монодия) напев конкретного дня повторялся в течение соответствующей ему по порядку недели. Восемь недель составляли столп, который циклически повторялся в течение года.

В русском осмогласии понятие гласа как лада деформировалось, а средством мелодической характеристики гласа стала выступать *совокупность попевок*, т. е. мелодических оборотов (в нашей терминологии — *элементарных структурных единиц* знаменного распева). Наиболее характерные для каждого гласа попевки представлены в подборке В. М. Металлова [5]. Но они не определяют однозначным образом гласовую принадлежность песнопения, поскольку некоторые попевки встречаются в песнопениях разных гласов.

Восемь гласов охватывают практически весь попевочный фонд церковного пения. Каждое песнопение, как правило, принадлежит одному из 8 гласов и строится на попевках данного гласа. Гласовая принадлежность песнопения обычно указывается в явном виде. Сказанное выше означает, что любой алгоритм дешифровки должен учитывать подчиненность песнопения системе осмогласия, поскольку одно и то же знамя может иметь разную интерпретацию в зависимости от гласа, типа структурной единицы, в состав которой оно входит, ее позиции в тексте и ряда других факторов.

Характер многозначности соответствия «знамя–нота» отражен в построенной авторами на основе двознаменников электронной азбуке знаменного распева [6]. Некоторые знамена, например такие, как  (статья закрытая малая) или (статья простая с подверткой), имеют до 10 различных интерпретаций, отличающихся друг от друга интервально-ритмическими характеристиками. При этом каждая из интерпретаций может иметь несколько звуковысотных привязок. Именно с *многозначностью соответствия «знамя–нота»* связаны основные проблемы дешифровки.

Начиная с XVII в., знамена начинают снабжать *степенными и указательными пометами*. Первые уточняют высоту распева знамени, вторые — особенности его распева. Термин «читаемые» относится к пометным песнопениям, хотя и здесь однозначный результат не гарантирован (см. [3, 4]). Беспометные рукописи XVI в. и более раннего периода практически нечитаемы.

Для устранения неоднозначности можно привлекать контекст, что эквивалентно переходу от отдельных знамен к более крупным структурным единицам, таким как *попевки, лица, фиты* и др. Однако известные подборки этих структурных единиц (азбуки, кокизники, фитники) малопригодны для целей дешифровки из-за формы их представления. Так, в подборке В. М. Металлова [5] попевки приведены только в нотолинейной форме, а в подборке М. В. Бражникова [7] около половины лиц и фит имеют лишь знаменное представление вместо двознаменного, требуемого для дешифровки.

Песнопения, представленные в двознаменниках, записаны в диапазоне *обиходного звукоряда* (рис. 1), включающего ноты G, A, H (малой октавы), c, d, e, f, g, a, b (первой октавы), C, D (второй октавы). В пометных рукописях ступеням звукоряда соответствуют обозначения  ,  ,  ,  ,  ,  ,  ,  ,  ,  ,  . Соответствующая знамени степенная помета указывает на наивысший звук в его распеве [1].

Особенности нотолинейной интерпретации знамен поясняются с помощью указательных помет: (или) — тихая; — ломка; — борзая; — ударка; — качка (или купно); — зевок; — равно. Знамена с указательной пометой и без нее трактуются как разные, поскольку могут иметь отличающиеся распевы. Длительности звуков обозначаются следующим образом:  — 1 (целая); — 2 (половинная); — 4 (четвертная); — 8 (восьмая). Для обозначения высоты и длительности звука используем комбинацию буквы и цифры (например, $H4$ — это четвертная нота «си» малой октавы). Знак $(*)$ используется в качестве разделителя между нотолинейными интерпретациями разных знамен,

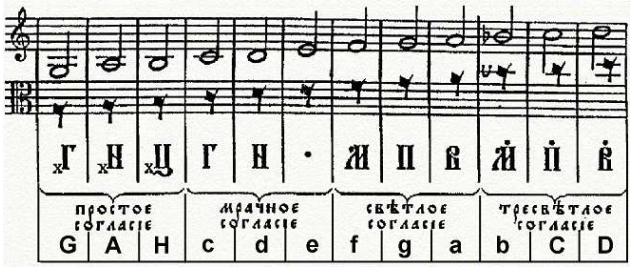


Рис. 1 Обиходный звукоряд

а ($\sim$) — как символ эквивалентности, отделяющий знаменную цепочку от ее нотолинейного представления (например, запись $\sim d4c4d2 * c4e4$ означает, что цепочка из двух знамен, стоящая слева от ($\sim$), интерпретируется в тексте двознаменника, соответственно, цепочками из трех и двух нот, отделенными друг от друга знаком ($*$)).

3 Описание подхода

На данный момент мы располагаем электронными версиями трех двознаменников конца XVII – начала XVIII в. Это разные редакции многожанровой певческой книги «Октоих» из Соловецкого собрания Российской национальной библиотеки, г. С.-Петербург (шифры 618/644, 619/647, QI188). Число песнопений в разных гласах двух первых Октоихов варьирует в диапазоне от 25 до 29, для QI188 этот показатель на треть ниже. Длины песнопений составляют от нескольких десятков до двух–трех сотен знамен. На основе этого (обучающего) материала была построена электронная азбука знаменного распева [6], выгодно отличающаяся от известных авторских азбук по многим показателям, в частности наличием информации о частоте встречаемости и звуковысотной привязке различных интерпретаций каждого знамени в каждом гласе.

Анализ электронной азбуки показал, что при всей вариативности знаменного распева в каждом гласе существует некоторое количество знамен, не меняющих своей интерпретации. Обычно их доля не превышает 10% от размера алфавита. В связи с этим возникло предположение о наличии в общем случае повторяющихся[1] цепочек знамен произвольной длины L ($L \geqslant 1$), однозначно или почти однозначно (с допустимыми отклонениями) интерпретируемых в пределах одного гласа. Цепочки первого типа естественно называть *внутригласовыми инвариантами* (ВИ), а второго — *квазиинвариантами* (КВИ). Их можно трактовать как особый тип структурных единиц, своего рода «островки стабильности», характеризующиеся существенным снижением уровня неоднозначности в интерпретации знамен.

В работах [8, 9] гипотеза о наличии в текстах двознаменников цепочек знамен со свойствами ВИ и КВИ подтверждена экспериментально на указанном выше материале. Для каждого из 8 гласов построены словари ВИ и КВИ, разбитые на подсловари в соответствии со значениями $L = 1, 2, \ldots$ Алгоритмы выделения ВИ и КВИ основаны на следующих соображениях:

1) элементами словарей ВИ и КВИ могут быть, как уже говорилось выше, лишь цепочки знамен, повторяющиеся в конкретном гласе, так как понятие инвариантности

[1]Требование повторяемости является принципиальным, поскольку оно исключает из рассмотрения однократно встречающиеся цепочки знамен. Такими, начиная с некоторого L, являются практически все достаточно длинные цепочки

подразумевает сохранение нотолинейной интерпретации (в строгом смысле или с допустимыми вариациями) при повторении цепочки. Требование повторяемости в неявной форме ограничивает длины ВИ и КВИ, так как с увеличением длин цепочек частота F их встречаемости в гласе падает. Число просматриваемых цепочек можно ограничить, введя пороговое значение F_r. При $F < F_r$ цепочка выбывает из рассмотрения. Будем использовать пороговое значение $F_r = 3$;

2) цепочка, составленная из многозначных знамен, может оказаться однозначно или почти однозначно интерпретируемой в гласе. Поэтому при формировании словарей ВИ и КВИ должен быть рассмотрен полный спектр цепочек произвольной длины с $F \geqslant F_r$ в подборке песнопений, представляющих конкретный глас;
3) формировать словари ВИ и КВИ следует на основе *беспометных* текстов. Тогда словари будут иметь формат «беспометное знамя – нота», единственно пригодный для решения задачи дешифровки в общем случае (беспометная нотация). Указанные двознаменники содержат пометы, которые следует предварительно устранить. Это равносильно агрегированию исходного знаменного алфавита, при котором в одну группу объединяются и обозначаются одним символом все пометные и беспометные варианты конкретного знамени. Так, в одну группу будут объединены  (стрела мрачная крыжевая с оттяжкой), она же с указательной пометой «борзо» () и «тихая» (), имеющие разные распевы. Предполагается, что в словарях ВИ и КВИ «агрегированная стрела» будет встречаться в разных контекстах, определяющих тот или иной вариант ее распева;
4) возможны различные варианты определения КВИ. В используемом авторами приближении цепочка знамен трактуется как КВИ, если одна из ее интерпретаций доминирует над остальными по частоте, а именно: выполняется соотношение $F_{\mathrm{dom}}/F > 1/2$, где F — частота встречаемости цепочки в песнопениях гласа; F_{dom} — максимальная из частот встречаемости ее интерпретаций. Доминирующая интерпретация трактуется в дальнейшем как значение КВИ. Важно отметить, что понятия ВИ и КВИ имеют относительный характер: они зависят от объема и состава исходной (обучающей) подборки. При увеличении ее объема или изменении жанрового состава отдельные ВИ могут перейти в категорию КВИ, а некоторые КВИ перестают быть таковыми («исчезает» доминирующая интерпретация). Эффекты такого рода можно проследить при переходе от отдельных Октоихов (так строились словари, описанные в [8, 9]) к их совокупности (в данной работе словари формируются на основе всех трех Октоихов).

Построенные по двознаменным Октоихам словари ВИ и КВИ, представленные в формате «беспометное знамя – нота», служат основой для нотолинейной реконструкции контрольных *беспометных* песнопений. В качестве контрольного материала авторы использовали пометный двознаменник «Праздники» (РНБ, г. С.-Петербург, Кирилло-Белозерское собрание, шифр 797/1054) и фрагмент двознаменного Ирмология из собрания В. Ф. Одоевского (М., РГБ, Ф210. №18). Пометы в обоих случаях предварительно удалялись, т. е. на вход алгоритма дешифровки подавались беспометные знаменные песнопения, а нотолинейная компонента указанных двознаменников использовалась лишь для сравнения реконструированного распева с реальным.

Схема дешифровки фактически сводится к покрытию песнопения всевозможными цепочками знамен из словарей ВИ и КВИ и приписыванию найденным фрагментам их певческого значения, зафиксированного в словарях. При этом возможны отказы (не существует ни одного ВИ или КВИ, покрывающего данное знамя) и конфликты интересов (знамя

входит в состав разных ВИ или КВИ, где ему приписываются неидентичные певческие значения). Более детально процесс дешифровки описан ниже.

4 Алгоритм дешифровки. Результаты апробации

Пусть $T = t_1 t_2 \cdots t_n \cdots t_N$ — представление песнопения одного из 8 гласов в виде последовательности знамен. Гласовая принадлежность песнопения обычно известна, хотя возможны исключения, и тогда это становится предметом отдельного рассмотрения. Покрытие песнопения цепочками знамен из словарей ВИ и КВИ, построенных для данного гласа, осуществляется итеративно по L, т. е. сначала используются ВИ и КВИ длины 1, затем длины 2 и т. д. При $L = 1$ произвольное знамя t_n песнопения, расположенное в n-й позиции, может быть покрыто единожды, если найдется идентичный ему ВИ или КВИ длины 1. В таком случае открываем список возможных интерпретаций знамени t_n из подсловаря инвариантов длины 1. При $L = 2$ знамя t_n может быть покрыто дважды, если среди ВИ (КВИ) длины 2 найдутся цепочки, совпадающие с биграммами $t_{n-1}t_n$ и $t_n t_{n+1}$ песнопения. В этом случае пополняем список возможных интерпретаций знамени t_n его значениями в составе указанных выше биграммных инвариантов. Знамя t_n при $L = 2$ может быть покрыто и единожды (заносим в список лишь одну интерпретацию) или не покрыто вовсе.

При $L \geqslant 3$ на каждой итерации имеем не более чем L возможностей покрытия знамени t_n цепочками ВИ и КВИ длины L. Продолжаем пополнять список возможных интерпретаций знамени t_n вплоть до исчерпания словарей ВИ и КВИ для данного гласа. Из множества полученных певческих значений для t_n выбираем самое частое. При равенстве голосов фиксируем случайным образом любой вариант.

По итогам описанной процедуры каждое знамя в песнопении либо получает единственную интерпретацию, хотя и не всегда правильную, либо не покрывается ни одной из цепочек словарей ВИ и КВИ. Сравнение результатов реконструкции с реальным нотолинейным текстом позволяет оценить количество правильно интерпретированных знамен (n_+) и число знамен, трактовка которых отличается от истинной (n_-). Коэффициент покрытия песнопения цепочками из словарей ВИ и КВИ определяем тогда в виде $k = n_+/N$, где N — число знамен в песнопении ($0 \leqslant k \leqslant 1$). Коэффициент покрытия гласа, представленного группой контрольных песнопений, можно определить либо как среднее коэффициентов покрытия отдельных песнопений, либо в виде отношения $\overline{k} = \overline{n_+}/\overline{N}$, где $\overline{n_+}$ — суммарное число знамен, правильно интерпретированных во всех песнопениях гласа, а $\overline{N}$ — суммарная длина песнопений. Показатель $\overline{k}$ характеризует эффективность использования словарей ВИ и КВИ для дешифровки беспометных песнопений.

Для апробации подхода было проведено несколько экспериментов. В первом из них разделение данных, используемых для обучения и контроля, осуществлялось с помощью процедуры «скользящего контроля». Суть ее (применительно к конкретному гласу) в следующем:

- песнопения i-го гласа ($1 \leqslant i \leqslant 8$) из трех Октоихов объединяются в одну группу $T_i = T_{i1}T_{i2}, \ldots$);
- удаляется песнопение T_{i1}, а по оставшимся строятся словари ВИ и КВИ;
- с помощью построенных словарей вычисляется коэффициент покрытия k_{i1} удаленного песнопения T_{i1};
- T_{i1} возвращается в исходную подборку, удаляется следующее песнопение T_{i2}, и процесс повторяется до исчерпания всех n_i песнопений i-й группы;

Таблица 1 Результаты первого эксперимента

Гласы	$\overline{k}$
1	0,786
2	0,701
3	0,744
4	0,730
5	0,735
6	0,685
7	0,801
8	0,725

Таблица 2 Результаты второго эксперимента

Гласы	$\overline{k}$
1	0,690
2	0,518
3	0,726
4	0,652
5	0,603
6	0,546
7	0,602
8	0,543

– вычисляется усредненный показатель покрываемости песнопений i-го гласа: $\overline{k_i} = (\sum_{l=1}^{n_i} k_{il})/n_i$.

Результаты эксперимента представлены в табл. 1. Из них следует, что в среднем по гласам правильно реконструируются порядка 70%–80% знамен. Однако эти оценки несколько завышены, поскольку поочередно удаляемые песнопения сохраняли какие-то связи с оставшимися, что при реконструкции повышало их шансы.

Более реальные данные получены в ходе второго эксперимента, когда для контроля использовался двознаменник «Праздники» из Кирилло-Белозерского собрания (другая школа). Его в основном составляют стихиры, но они значительно отличаются от стихир из Октоихов сложностью и развитостью распева, вариативностью, широтой динамического диапазона. Количество песнопений в гласах было сопоставимо с аналогичным показателем для Октоихов 619/647 и 618/644 за исключением гласов 3 и 7, слабо представленных в двознаменнике «Праздники». Использовался второй вариант коэффициента покрытия ($\overline{k} = \overline{n_+}/\overline{N}$). Результаты этого эксперимента приведены в табл. 2.

Показатели покрываемости заметно снижаются, что объясняется существенными различиями в обучающем и контрольном материале.

В третьем эксперименте контрольный материал был представлен 22 песнопениями первого гласа Ирмология. Коэффициент покрываемости этих песнопений инвариантами первого гласа, полученными на материале трех Октоихов, составил 0,687.

5 Обсуждение результатов

1. Словари ВИ и КВИ, построенные на основе трех Октоихов, достаточно представительны (в сумме порядка полутора тысяч цепочек знамен разной длины в каждом гласе). Они в состоянии обеспечить дешифруемость контрольного материала, близкую в среднем к 60%–70% даже при отсутствии помет у знамен. С определенной осторожностью подход может быть применен и к беспометным песнопениям XVI в. У авторов имеется минимальный опыт такого рода, однако достоверность реконструкции требует дополнительного исследования из-за отсутствия двознаменников, датированных указанным периодом.
2. Анализ покрываемости отдельных песнопений выявил редко встречающиеся аномалии — песнопения с коэффициентами покрытия порядка 0,2 и ниже. Предварительный анализ показал, что причинами могут быть: *неверно указанная гласовая принадлежность* песнопения (оно плохо дешифруется по словарю «собственного» гласа, но гораздо лучше по словарю другого); *неоговариваемые звуковысотные переносы* отдельных

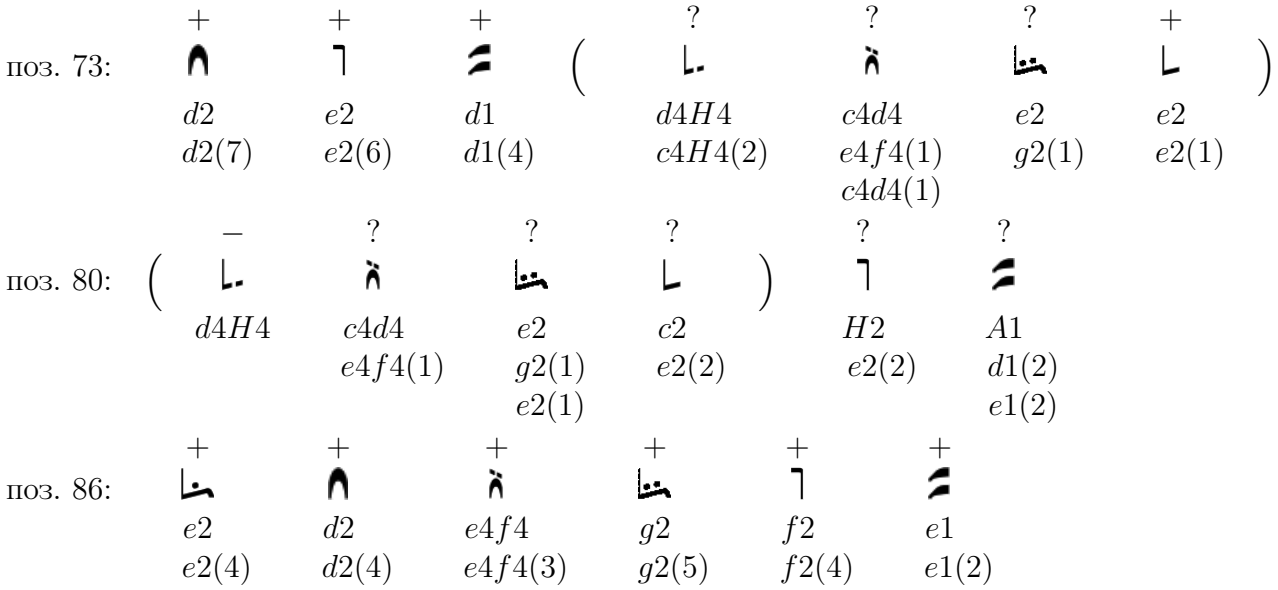


Рис. 2 Пример дешифровки фрагмента песнопения

фрагментов мелодии; *наличие в песнопении большого количества лиц и фит* — структурных единиц, относящихся к категории мелизматических украшений мелодии, когда на один слог текста приходится много звуков мелодии. Разводы лиц и фит «рядовым» знаменем не всегда стандартны. Их можно сравнить с идиомами в естественном языке. Индикатором лиц и фит в распеве является стихотворная компонента параллельных текстов.

3. Ошибочно трактуемые знамена часто располагаются между попевочными структурами. Среди таких «связующих» знамен много высокочастотных: ,  ,  ,  и пр. Порой они образуют тандемные повторы. Это удобный способ разнесения основных структурных единиц текста (в данном случае попевок) на нужное расстояние, используемый в разных языковых системах. Сложность дешифровки тандемов в том, что тандемному повтору на знаменном уровне не всегда соответствует повтор на нотолинейном уровне (и наоборот). Приведем реальный пример (рис. 2) дешифровки фрагмента песнопения первого гласа из двознаменника «Праздники» (лист 78 оборотный). Фрагмент начинается с 73-й позиции.

   Здесь каждая строка двознаменного текста — это последовательность знамен, а ниже — их певческие значения. Каждое знамя помечено сверху одним из трех символов: «–» означает, что знамя не покрыто ни одним ВИ или КВИ; «+» означает, что знамя интерпретировано правильно; в третьей строке его певческое значение повторено с указанием (в скобках) числа проголосовавших за него ВИ и КВИ; «?» говорит о том, что знамя интерпретировано неверно, в этом случае в третьей строке приводится ошибочная трактовка, одна или несколько (строки 4, 5 и т. д.), если за них подано одинаковое число голосов. В строках 4, 5 и т. д. может быть представлено и истинное значение (см., например, знамена 77, 82), но механизм случайного выбора, упрощенный до предела, сработал не в их пользу (выбирался первый по списку кандидат среди равных по числу голосов).

   Пример не слишком богат попевочными структурами: их три, и все они, по классификации А. Н. Кручининой [10], являются разновидностями «кокиз», кадансовая структура которых представлена позициями 73–75, 83–85, 89–91. Наибольший же интерес

представляет фрагмент между первым и вторым вхождением «кокиз» (поз. 76–83), содержащий тандемный повтор из 4 знамен (выделен скобками). Нетрудно видеть, что на нотолинейном уровне повтор искажается при переходе от «крюка светлого» () к стопице (). Практически весь указанный фрагмент характеризуется пониженной дешифруемостью, хотя отличия от реального текста (вторая строка) касаются, в основном, лишь звуковысотной привязки (интервально-ритмическая структура не нарушается). Более того, интерпретация «стопицы с очком» (, поз. 76) в тексте двузнаменника (скачок вниз на две ступени) представляется менее предпочтительной, чем стандартный ход на одну ступень в реконструированном тексте.

Приведенный пример показывает, что тандемные повторы претендуют на роль самостоятельных структурных единиц знаменного распева и нуждаются в отдельном изучении. Первый шаг в этом направлении сделан в [11]. Следует отметить также, что проявления тандемной повторности в знаменном распеве зависят от жанровой специфики. Так, например, биграмма  часто используется в тандемном варианте (  $\sim e2{*}d1{*}e2{*}d1$) в песнопениях гласа 4 («Праздники»), а в том же гласе всех Октоихов встречается всего 4 раза, не образует тандемных вхождений и имеет доминирующую интерпретацию $c2{*}H1$, отличную от $e2{*}d1$. Как следствие, данная цепочка в песнопениях гласа 4 («Праздники») интерпретируется неверно. Из этого следует, что кроме настройки на конкретный глас желательна настройка и на конкретный жанр.

4. Созданные для дешифровки электронные словари ВИ и КВИ можно рассматривать как *систему описания гласов*, каждый из которых характеризуется достаточно представительным набором песнопений. Эти словари могут быть использованы не только для дешифровки песнопений, но и для решения вспомогательных задач, таких как определение гласовой принадлежности песнопения, обнаружение ошибок кодирования, выявление функционирующих в гласе структурных единиц и устойчивых их комбинаций, а также для *количественной и качественной характеризации системы осмогласия* в целом. В последнем случае речь идет о выявлении сходства и различий между гласами.

Применительно к задаче дешифровки нас больше интересует сходство гласов, а именно: наличие достаточно длинных общих для разных гласов цепочек знамен, одинаково интерпретируемых в разных гласах — своего рода *межгласовых инвариантов* (МИ). Степень дешифруемости таких цепочек в разных гласах с помощью словарей ВИ и КВИ может сильно отличаться, поскольку она зависит от частоты встречаемости цепочки в конкретном гласе и ее вариативности. В одном гласе она может попасть в словари ВИ и КВИ и иметь хорошие показатели по дешифруемости, а в другом — наоборот. В таком случае возникает возможность использования взаимосвязей между гласами на уровне МИ для улучшения показателей дешифруемости одного гласа за счет другого. Фактически речь идет о дешифровке *по прецедентам* (порой однократно встречающимся). Удобным инструментом для выявления межгласовых связей на уровне МИ является предложенный авторами аппарат сложностных разложений [12], в основе которого применительно к рассматриваемому случаю лежит представление песнопений одного гласа в виде конкатенации цепочек знамен (или нот) из песнопений другого гласа.

Цепочек, которые можно отнести к категории МИ, достаточно много. Приведем лишь один пример, иллюстрирующий возможность дешифровки по прецедентам. При сравнении гласов 1 и 4 Октоиха 619/647 выявляется общая цепочка знамен из разных песнопений (глас 1: «Слава и ныне. Богородичен», поз. 45 и глас 4: «На стиховне стихира», поз. 63):

$f4e4d2 \quad f4g4 \quad f8e8d4 \quad f4g2f4 \quad e1 \quad d1.$

В соответствии с [10] — это одна из разновидностей «кулизм». С помощью словарей ВИ и КВИ, построенных для данного Октоиха, цепочка из гласа 1 дешифруется правильно, а из гласа 4 — с ошибкой в интерпретации второго и третьего знамени. Эта ошибка может быть скорректирована на основе информации о наличии указанного выше МИ в обучающей подборке.

6 Заключение

Проблема нотолинейной реконструкции (дешифровки) древнерусских церковных песнопений, представленных в знаменной форме записи, является одной из наиболее известных и актуальных в музыкальной медиевистике. Трудности ее решения связаны с многозначностью соответствия «знамя–нота». Авторы развивают новый компьютерно-ориентированный подход к ее решению, основанный на использовании двознаменников конца XVII — начала XVIII в., записанных в виде трех *параллельных* (синхронизованных между собой) *текстов*: знаменного, нотолинейного и стихотворного (старославянского). Двознаменники нужны для того, чтобы извлечь из них информацию о наиболее устойчивых (наименее вариативных) структурных единицах знаменного распева, названных авторами инвариантами и квазиинвариантами. Последние представлены в единственно пригодном для дешифровки формате «беспометное знамя – нота».

На достаточно объемном обучающем материале построены электронные словари инвариантов и квазиинвариантов для каждого из гласов знаменного распева. Разработан алгоритм дешифровки знаменной нотации с использованием указанных словарей. Эксперименты на независимом контрольном материале показали, что даже без апелляции к структурной организации знаменного распева (резерв для дальнейшего развития подхода) удается обеспечить в среднем 60%–70%-ную дешифруемость по гласам.

Разбор ошибок говорит о том, что часть из них — это ошибки кодирования (авторские) и погрешности в самой певческой книге. Другая часть ошибок носит характер допустимого варьирования. Случаев радикального искажения нотолинейной структуры не отмечено.

Развитие подхода мыслится в направлении привлечения информации о внутренней структуре песнопений (попевки, лица, фиты, тандемные повторы). Важной может оказаться информация о взаимосвязях между гласами (приведены некоторые примеры и соображения на указанную тему). И, наконец, ощутимую пользу может принести настройка на отдельные классы песнопений (стихиры, ирмосы и др.).

К основным достоинствам подхода можно отнести опору на *двознаменники* периода наивысшего расцвета знаменного пения, ориентацию на *беспометную нотацию* и экспериментально подтвержденную применимость к песнопениям разного жанра.

Литература

[1] *Бражников М. В.* Древнерусская теория музыки. — Л.: Музыка, 1972. 423 с.

[2] *Кутузов Б. П.* Русское знаменное пение. — 2-е изд. — М., 2008. 304 с.

[3] *Бахмутова И. В., Гусев В. Д., Титкова Т. Н.* О функциях указательных помет (на материале двознаменника XVIII века) // Сибирский музыкальный альманах. — Новосибирск: Изд-во НГК, 2002. С. 81–92.

[4] *Бахмутова И. В., Гусев В. Д., Титкова Т. Н.* Факторы, влияющие на точность нотолинейной реконструкции пометных знаменных песнопений // Сибирский музыкальный альманах. — Новосибирск: Изд–во НГК, 2004. С. 51–59.

[5] *Металлов В. М.* Осмогласие знаменного распева (сборник нотолинейных попевок). — М., 1899, 50 с.

[6] *Бахмутова И. В., Гусев В. Д., Титкова Т. Н.* Электронная азбука знаменного распева: Предварительная версия // Вычислительные системы, 2005. Вып. 174. С. 29–53.

[7] *Бражников М.В.* Лица и фиты знаменного распева. — Л.: Музыка, 1984. 302 с.

[8] *Бахмутова И. В., Гусев В. Д., Титкова Т. Н.* Компьютерный поиск инвариантных структурных единиц знаменного распева // Проблемы музыкальной науки. Российский научный специализированный журнал, 2011. №1(8). С. 20–24.

[9] *Бахмутова И. В., Гусев В. Д., Титкова Т. Н.* Выявление инвариантов и квазиинвариантов знаменного распева с помощью билингв типа «знамя-нота» // Мат-лы Всеросс. конф. ЗОНТ–2013, 2013. Т. 1. С. 27–35.

[10] *Кручинина А. Н.* Попевка в русской музыкальной теории XVII века. Автореф. дисс. . . . канд. иск. н. — Л., 1979.

[11] *Бахмутова И. В., Гусев В. Д., Мирошниченко Л. А., Титкова Т. Н.* Тандемные повторы в знаменных песнопениях // Вычислительные системы, 2005. Вып. 174. С. 13–28.

[12] *Gusev V. D., Miroshnichenko L. A.* Complexity expansions in comparisons of character sequences // Pattern Recogn. Image Anal., 2014. Vol. 24. Iss. 4. P. 467–472. http://link.springer.com/article/10.1134/S1054661814040063.

Поступила в редакцию 15.06.2015

References

[1] Brazhnikov, M. V. 1972. *Drevnerusskaya teoriya muzyki* [Ancient Russian theory of music]. Leningrad: Music. 423 p.

[2] Kutuzov, B. P. 2008. *Russkoe znamennoe penie* [Russian znamennyi chant]. Moscow. 304 p.

[3] Bakhmutova, I. V., V. D. Gusev, and T N. Titkova. 2002. On functions of the indicative marks (in the material of dvoyeznamenniks XVIII sc.). *Siberian Musical Almanac*. Novosibirsk: NGK. 81–92.

[4] Bakhmutova, I. V., V. D. Gusev, and T. N. Titkova. 2004. Factors acting on accuracy of the noted reconstruction of Russian znamennyi marked hymnals. *Siberian Musical Almanac*. Novosibirsk: NGK. 51–59.

[5] Metallov, V. M. 1899. *Osmoglasie of znamennyi chant* [Set of noted popevoks]. Moscow. 50 p.

[6] Bakhmutova, I. V., V. D. Gusev, and T N. Titkova. 2005. Electronic alphabeth of Russian chant: Previous version. *Computing Syst.* 174:29–53.

[7] Brazhnikov, M. V. 1984. *Litsa and fitas of Russian znamennyi chant*. Leningrad: Music. 302 p.

[8] Bakhmutova, I. V., V. D. Gusev, and T N. Titkova. 2011. Computerized search for the invariant structural elements of znamennyi chant. *Music Scholarship* 1(8):20–24.

[9] Bakhmutova, I. V., V. D. Gusev, and T N. Titkova. 2013. Revelation of invariants and quasi-invariants of znamennyi chant using “neume-note” bilinguas. *Russian Conference KONT-2013 Proceedings*. Novosibirsk. 1:27–35.

[10] Kruchinina, A. N. 1979. Popevka in Russian musical theory of XVII sc. Ph.D. Thesis. Leningrad.

[11] Bakhmutova, I. V., V. D. Gusev, L. A. Miroshnichenko, and T. N. Titkova. 2005. Tandem repeats in the neume hymns. *Computing Systems* 174:13–28.

[12] Gusev, V. D., and L. A. Miroshnichenko. 2014. Complexity expansions in comparisons of character sequences. *Pattern Recogn. Image Anal.* 24(4):467–472. Available at: http://link.springer.com/article/10.1134/S1054661814040063 (accessed December 28, 2015).

Received June 15, 2015

Relevance tagging machine*

***D. A. Molchanov*[1], *D. A. Kondrashkin*[2], *and D. P. Vetrov*[2]**
dmolch111@gmail.com; kondra2lp@gmail.com; vetrovd@yandex.ru
[1]Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, Russia
[2]National Research University Higher School of Economics, 20 Myasnitskaya str., Moscow, Russia

In many classification or regression problems, there may be a lot of irrelevant features. Bayesian automatic relevance determination (ARD) is a popular approach to feature selection. However, the application area of this approach has been limited. In this paper, this approach is utilized in a more general case and it is applied to a binary classification problem with binary features. Also, a new binary classification model and a learning algorithm that can purge unwanted features from the model have been developed.

Keywords: *binary classification; feature selection; automatic relevance determination; sparse bayesian learning; variational lower bounds*

DOI: 10.21469/22233792.1.13.09

Машина релевантных тегов*

***Д. А. Молчанов*[1], *Д. А. Кондрашкин*[2], *Д. П. Ветров*[2]**
[1]Московский государственный университет им. М. В. Ломоносова, Москва, Россия
[2]Национальный исследовательский университет «Высшая школа экономики», Москва, Россия

При решении многих задач классификации или регрессии зачастую приходится сталкиваться с большим количеством нерелевантных признаков. Одним из известных способов решения задачи отбора признаков является метод, основанный на Байесовском подходе к выбору модели. Этот метод получил широкое распространение, однако область его применения была ограничена. В данной работе этот метод применяется для более широкого класса моделей и исследуется на примере задачи бинарной классификации с бинарными признаками. Также предложена новая модель для бинарной классификации данных и метод обучения этой модели, позволяющий автоматически убирать нерелевантные признаки.

Ключевые слова: *бинарная классификация; отбор признаков; автоматическое определение релевантности; вариационные нижние оценки*

DOI: 10.21469/22233792.1.13.09

1 Introduction

Feature selection is an important challenge that arises in most machine learning problems. There are different approaches to this task. One of them is to use predictive models that can automatically choose the most relevant features during the training procedure. For example, it can be done with LASSO (Least Absolute Shrinkage and Selection Operator) regression or other models that use L1-regularization to ensure sparsity. Bayesian ARD [1]) is another approach to

*This research is funded by RFBR grant #15-31-20596mol-a-ved, Microsoft Research, research initiative: Computer vision collaborative research in Russia, Skoltech SDP Initiative, applications A1 and A2.

developing such models. As an example, consider the Relevance Vector Machine (RVM), [2]. In case of regression, the RVM is a linear model with an ARD prior; $\boldsymbol{x}$ is an object; t is its target; $\boldsymbol{w}$ is a vector of model parameters or weights; and $\boldsymbol{\varphi}(\boldsymbol{x})$ is a vector of generalized features. The model definition is shown below:

$$\boldsymbol{w} = (w_1, \ldots, w_M)^{\mathrm{T}} \in \mathbb{R}^M; \quad \boldsymbol{\varphi}(\boldsymbol{x}) = (\varphi_1(\boldsymbol{x}), \ldots, \varphi_M(\boldsymbol{x}))^T \in \mathbb{R}^M, \quad t \in \mathbb{R};$$

$$y(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\varphi}(\boldsymbol{x}); \tag{1}$$

$$p(t \,|\, \boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(t \,|\, y(\boldsymbol{x}), \beta^{-1}); \tag{2}$$

$$p(\boldsymbol{w} \,|\, \boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i \,|\, 0, \alpha_i^{-1}) \tag{3}$$

and the following expression is the marginal likelihood function, also known as evidence [3]:

$$p(\boldsymbol{t} \,|\, \boldsymbol{X}, \boldsymbol{\alpha}, \beta) = \int p(\boldsymbol{t} \,|\, \boldsymbol{X}, \boldsymbol{w}, \beta) p(\boldsymbol{w} \,|\, \boldsymbol{\alpha})\, d\boldsymbol{w}\,.$$

Equation (2) defines the likelihood function for an object $\boldsymbol{x}$. Here, β is the noise precision, $\beta = \sigma^{-2}$, and $y(\boldsymbol{x})$ is the mean of the target function given by a linear model defined in (1). Expression (3) describes the prior over the weight parameters $\boldsymbol{w}$ (ARD prior). When the evidence of the model is maximized with respect to hyperparameters $\boldsymbol{\alpha}$, some of them go to infinity. The corresponding weight parameters will then have posterior distributions that are concentrated at zero; so, the corresponding basis functions $\varphi_i(\boldsymbol{x})$ are pruned out of the model. This effect is known as ARD effect and is explained and discussed in [1, 2] and [4, p. 349–353].

However, this effect is usually studied on models with Gaussian prior. The present authors propose to extend this approach and use another family of distributions. In this paper, a binary classification problem with binary features is considered as an example. A new probabilistic model has been developed for this task and beta prior distribution has been used to reproduce ARD effect.

2 Model of Relevance Tagging Machine

2.1 Probabilistic model

Consider a binary classification problem of objects that have binary features (tags). Let $(\boldsymbol{x}_i, t_i)_{i=1}^n$ be the training set, where $\boldsymbol{x}_i$ is the object described by a binary vector, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^{\mathrm{T}}$, d denotes the number of tags, and $t_i \in \{0, 1\}$ is the class label. In this notation, $x_{ij} = 1$ if object $\boldsymbol{x}_i$ has tag j and $x_{ij} = 0$ otherwise.

Under the assumption that all tags affect the class label independently, we define the probabilistic model of relevance tagging machine (RTM):

$$\mathsf{P}(t = 1 \,|\, x_j = 1) = q_j;$$

$$\mathsf{P}(t = 1 \,|\, \boldsymbol{x}, \boldsymbol{q}) = \prod_{j=1}^{d} q_j^{x_j} \times \left(\prod_{j=1}^{d} q_j^{x_j} + \prod_{j=1}^{d} (1 - q_j)^{x_j} \right)^{-1}$$

where $\boldsymbol{q} = (q_1, \ldots, q_d)^{\mathrm{T}}$ are the model parameters, which are responsible for the tags' influence on the class label.

2.2 Bayesian Automatic Relevance Determination approach

Similarly to the RVM, follow a traditional Bayesian ARD approach to feature selection. The basic idea is to treat parameters $\boldsymbol{q}$ as random variables and place independent priors over them. As the domain of q_j is $[0, 1]$, it is natural to use beta distribution over q_j. Also, as both classes are meant to be of the same importance, symmetrical distribution is used:

$$q_j \sim \mathrm{Beta}(\alpha_j + 1, \alpha_j + 1),\ \alpha_j \in [0, +\infty).$$

Here, $\alpha_j = 0$ corresponds to the uniform distribution over q_j, so that there is no regularization of q_j. Contrary, if α_j tends to plus infinity, the variance of q_j tends to zero and that implies $q_j = 0.5$. It means that the jth tag is removed from the model:

$$\mathsf{P}(t \mid \boldsymbol{x}, \boldsymbol{q} = (q_1, \ldots, q_{j-1}, q_j = 0.5, q_{j+1}, \ldots, q_d)^{\mathrm{T}}) = \mathsf{P}(t \mid \boldsymbol{x}, \boldsymbol{q} = (q_1, \ldots, q_{j-1}, q_{j+1}, \ldots, q_d)^{\mathrm{T}}) = \mathsf{P}(t \mid \boldsymbol{x}, \boldsymbol{q}^{\setminus j}).$$

Note that the case $\alpha_j \in (-1, 0)$ is not considered because in this case, maximum a posteriori (MAP) estimate of $\boldsymbol{q}$ would be more contrast (i.e. closer to 0 or 1) than maximum likelihood (ML) estimate. In case of the problem under investigation, it is unreasonable to believe that a tag is actually more relevant than it seems to be.

The posterior is written using Bayes' theorem:

$$\mathsf{P}(\boldsymbol{q} \mid \boldsymbol{X}, \boldsymbol{t}, \boldsymbol{\alpha}) = \frac{\mathsf{P}(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{q}) p(\boldsymbol{q} \mid \boldsymbol{\alpha})}{\int \mathsf{P}(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{q}) p(\boldsymbol{q} \mid \boldsymbol{\alpha}) d\boldsymbol{q}}\,. \tag{4}$$

2.3 Evidence maximization

The denominator in Eq. (4) is called the *evidence* [3] of the model. In general, a simple model has higher evidence than the complex one if they have the same prediction accuracy [4, p. 349–352]. In the presented case, evidence maximization is expected to set $\alpha_j = +\infty$ for the majority of irrelevant features:

$$E(\boldsymbol{\alpha}) = \int \mathsf{P}(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{q}) p(\boldsymbol{q} \mid \boldsymbol{\alpha})\, d\boldsymbol{q} \to \max_{\boldsymbol{\alpha}}\,.$$

However, in the described model, likelihood and prior are not the conjugate distributions; so, the evidence is intractable. It also cannot be efficiently estimated numerically, because numerical computation of multidimensional integrals is a very difficult and time-consuming task. Therefore, one needs some kind of approximation in order to maximize the evidence. In this paper, an approach that uses variational lower bounds for optimization is described.

3 Variational lower bounds for evidence maximization

Definition 1. *A variational lower bound on a function* $f(\boldsymbol{w})$, $\boldsymbol{w} \in M \subseteq \mathbb{R}^n$, *is a function* $g(\boldsymbol{w}, \boldsymbol{\xi})$, $\boldsymbol{w} \in M$, $\boldsymbol{\xi} \in M$, *with the following properties:*

$$g(\boldsymbol{w}, \boldsymbol{w}) = f(\boldsymbol{w})\ \forall \boldsymbol{w} \in M;$$
$$g(\boldsymbol{w}, \boldsymbol{\xi}) \leqslant f(\boldsymbol{w})\ \forall \boldsymbol{w} \in M,\ \forall \boldsymbol{\xi} \in M,$$

$\boldsymbol{\xi}$ *is called a variational parameter.*

Consider an optimization problem $f(\boldsymbol{w}) \to \max\limits_{\boldsymbol{w}}$. If $g(\boldsymbol{w}, \boldsymbol{\xi})$ is a variational lower bound on $f(\boldsymbol{w})$, then this optimization problem can be solved in such coordinatewise optimization procedure:

$$\boldsymbol{w}^{k+1} = \arg\max_{\boldsymbol{w}} g(\boldsymbol{w}, \boldsymbol{\xi}^k); \quad \boldsymbol{\xi}^{k+1} = \boldsymbol{w}^{k+1}.$$

This optimization procedure is known as bound optimization algorithm or bound optimizer. Many popular optimization methods in machine learning and pattern recognition are the special cases of this algorithm. For instance, EM (expectation–maximization) algorithm and its extensions, generalized iterative scaling algorithm for maximum entropy models, nonnegative matrix factorization algorithm, and concave-convex procedure are the common examples of bound optimizers [5].

In general case, this lower bound may not be exact for any point $\boldsymbol{w}$ and variational parameters may be from a different space. In that case, the result of a similar optimization procedure

$$\boldsymbol{\xi}^{k+1} = \arg\max_{\boldsymbol{\xi}} g(\boldsymbol{w}^k, \boldsymbol{\xi}); \quad \boldsymbol{w}^{k+1} = \arg\max_{\boldsymbol{w}} g(\boldsymbol{w}, \boldsymbol{\xi}^{k+1}) \tag{5}$$

can be treated as an approximate solution of the original optimization task.

This approach is widely used in various optimization problems. The best feature of it is that there is no need to compute the original function $f(\boldsymbol{w})$. For example, a similar approach is used in [6] where it is applied to Bayesian logistic regression.

In case of RTM, this approach is applied to evidence maximization. A variational lower bound has been obtained on the evidence integrand and its integral has been used as a set of evidence lower bounds which can be used in an optimization procedure shown in (5).

Theorem 1. *Function* $\tilde{E}(\boldsymbol{\alpha}, \boldsymbol{H})$ *is a lower bound on RTM evidence for all* $\boldsymbol{H} \in (0,1)^{n\times d}$, $\boldsymbol{\alpha} \in [0, +\infty)^d$:

$$E(\boldsymbol{\alpha}) \geqslant \tilde{E}(\boldsymbol{\alpha}, \boldsymbol{H}) = \int \prod_{i=1}^{n} L_i(\boldsymbol{q}, \boldsymbol{\eta}_i) \prod_{j=1}^{d} p(q_j \mid \alpha_j)\, d\boldsymbol{q} =$$

$$= \left(\prod_{i=1}^{n} c_i(\boldsymbol{\eta}_i)\right) \prod_{j=1}^{d} \int \exp\left(\sum_{i:j\in Q_i} \tilde{c}_{ij}(\boldsymbol{\eta}_i) \left(\frac{1-q_j}{q_j}\right)^{|Q_i|(2t_i-1)}\right) p(q_j \mid \alpha_j)\, dq_j$$

$$\forall \boldsymbol{H} \in (0,1)^{n\times d},\ \forall \boldsymbol{\alpha} \in [0, +\infty)^d,$$

where

$$Q_i = \{j | x_{ij} = 1\};$$

$$c_i(\boldsymbol{\eta}_i) = \frac{\prod\limits_{j\in Q_i} \eta_{ij}^{t_i}(1-\eta_{ij})^{1-t_i}}{\prod\limits_{j\in Q_i} \eta_{ij} + \prod\limits_{j\in Q_i} (1-\eta_{ij})} \exp\left(\frac{\prod\limits_{j\in Q_i} \eta_{ij}^{1-t_i}(1-\eta_{ij})^{t_i}}{\prod\limits_{j\in Q_i} \eta_{ij} + \prod\limits_{j\in Q_i} (1-\eta_{ij})}\right);$$

$$\tilde{c}_{ij}(\boldsymbol{\eta}_i) = -\frac{\prod\limits_{j\in Q_i} \eta_{ij}^{t_i}(1-\eta_{ij})^{1-t_i}}{\prod\limits_{j\in Q_i} \eta_{ij} + \prod\limits_{j\in Q_i} (1-\eta_{ij})} \left(\frac{\eta_{ij}}{1-\eta_{ij}}\right)^{|Q_i|(2t_i-1)} |Q_i|^{-1};$$

and $\boldsymbol{H}$ *is the matrix of variational parameters and its* i*th row is equal to* $\boldsymbol{\eta}_i^{\mathrm{T}}$.

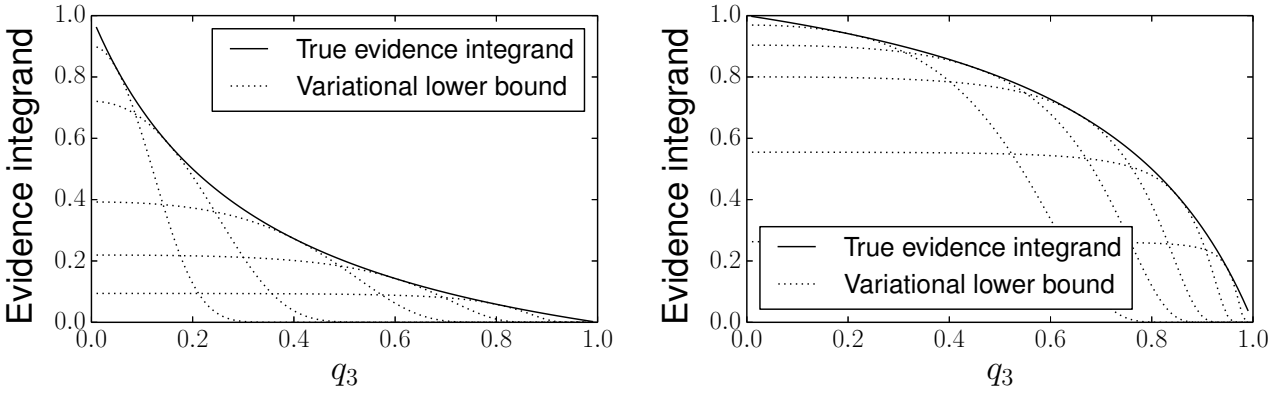


Figure 1 Evidence integrand variational lower bounds for a single object for different values of variational parameters η: (a) $t = 0$, $\boldsymbol{x} = (1, 0, 1)^{\mathrm{T}}$ and $q_1 = 0.8$; and (b) $t = 0$, $\boldsymbol{x} = (0, 1, 1)^{\mathrm{T}}$ and $q_2 = 0.2$

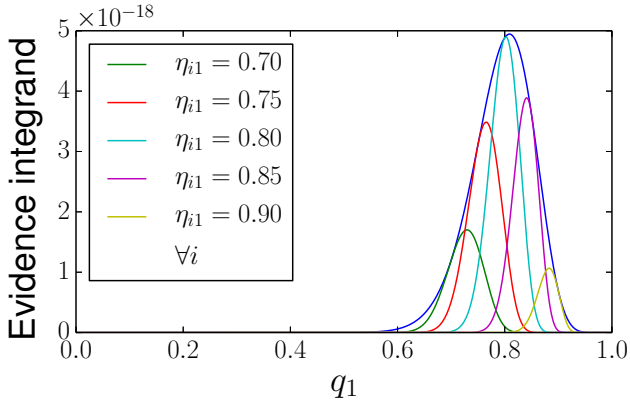


Figure 2 Evidence integrand lower bounds for the whole dataset

The proof of Theorem 1 is provided in Appendix A.

The shapes of the evidence integrand and its lower bounds are shown in Figs. 1 and 2. Note that although the evidence lower bound can be computed as a product of d one-dimensional integrals, these integrals still have to be computed numerically. Also, note that although the number of variational parameters is nd, usually most of them are inessential and do not affect the value of this lower bound. A variational parameter η_{ij} is essential if and only if $x_{ij} = 1$. Therefore, there are only $n\tau$ essential variational parameters where τ is the average number of tags per object.

The evidence lower bound is optimized in an EM-like algorithm:

1) E-step: $\boldsymbol{H}^{\mathrm{new}} = \arg\max_{\boldsymbol{H}} \log \tilde{E}(\boldsymbol{\alpha}^{\mathrm{old}}, \boldsymbol{H})$; and
2) M-step: $\boldsymbol{\alpha}^{\mathrm{new}} = \arg\max_{\boldsymbol{\alpha}} \log \tilde{E}(\boldsymbol{\alpha}, \boldsymbol{H}^{\mathrm{new}})$.

   These two steps are repeated until convergence.

On E-step, hyperparameters $\boldsymbol{\alpha}$ are fixed and variational parameters $\boldsymbol{H}$ are tuned to obtain the most accurate lower bound. On M-step, the best lower bound from the E-step is optimized

Table 1 Relevance determination performance on synthetic data

Noise	RTM-MAP-EM	RTM-EM	RVM	L1-LR
Percentage of removed irrelevant tags				
Random	99.64%	99.46%	99.10%	85.63%
Correlated	84.44%	88.90%	84.65%	100.00%
Percentage of removed genuine tags				
Random	4.50%	4.68%	2.63%	3.54%
Correlated	2.50%	4.34%	1.04%	2.50%

with respect to hyperparameters $\boldsymbol{\alpha}$. The L-BFGS-B method [8] was used to handle optimization problems on both steps of algorithm. This method is called RTM-EM.

Let k_E and k_M be the number of iterations of L-BFGS-B on E-step and M-step, respectively. The complexity of one iteration is then equal to $O(n\tau k_E + dk_M)$ operations of numerical integration.

Complexity of RTM-EM is too high; so, a simplification is suggested. In RTM-EM, on E-step of EM algorithm, an attempt to obtain the best possible value of variational parameters was made. Instead of that, a variational lower bound on the evidence integrand was used that is exact at its point of maximum. E-step will then look like this:

$$\boldsymbol{\eta}_i^{\text{new}} = \boldsymbol{q}^{\text{MAP}} = \arg\max_{\boldsymbol{q}} \mathsf{P}(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{q}) p(\boldsymbol{q} \mid \boldsymbol{\alpha}^{\text{old}}) \; \forall i.$$

Note that all objects share the same set of variational parameters; so, there are only d of them: $\boldsymbol{\eta}_i = \boldsymbol{\eta}_k$ for all $i, k = 1, \ldots, n$. This method was named RTM-MAP-EM. Its complexity is $O(dk_M)$ operations of numerical integration.

It was experimentally shown that RTM-MAP-EM also purges irrelevant (both noisy and correlated) tags and has comparable accuracy with RTM-EM algorithm. Also, both EM-based methods remove the majority of irrelevant tags on early steps. It means that only several steps of EM-algorithm are needed for feature selection. Further optimization will just tune the remaining hyperparameters.

The complete algorithm of RTM-MAP-EM is provided in Appendix B.

4 Experiments

4.1 Synthetic data

The ability of the presented methods to remove irrelevant features on a synthetic dataset was studied and compared to two classic feature selection models — RVM, where a similar idea is applied to linear regression, and L1-regularized logistic regression. There were 500 objects and 50 features. The data consisted of genuine tags that were used to generate the class label, and two types of irrelevant tags: random tags and tags that were correlated to some of the genuine tags. Relevance determination accuracy is shown in Table 1.

Both EM-based methods successfully remove nearly all random features and most correlated features. The RTM-MAP-EM, RTM-EM, and RVM give comparable results and detect random features better than L1-regularized logistic regression (L1–LR). However, L1–LR provided the best results in removing correlated tags.

4.2 Sentiment analysis

Also, the methods were tested on a real task: sentiment analysis problem [9]. In this problem, objects are sentences and the task is to classify them into positive and negative ones. A bag of

Table 2 Prediction performance on sentiment analysis dataset

Method	Prediction accuracy
RTM-MAP-EM	0.9659
RVM	0.9586
L1-LR	0.9708
RF	0.9416
GBDT	0.9683
SVM	0.9683

words representation was used (each tag represents a word; $x_{ij} = 1$ if object $\boldsymbol{x}_i$ contains the jth word from the dictionary). There were 1000 train objects, 411 test objects, and 1869 features. There were 11 tags per objects in average. Test set classification accuracy is shown in Table 2.

The present method (RTM-MAP-EM) was compared to different state-of-the-art classifiers like the RVM, L1-LR, Random Forest (RF), gradient boosting over decision stumps (GBDT), and SVM.

The present method provides prediction accuracy that is comparable to classical methods. It also provides a way to sort features with respect to their importance: RTM-MAP-EM chose about 70 tags to be relevant and removed everything else; L1-LR chose about 120 tags; and the RVM chose about 230 tags. A histogram of weights of most relevant words for these methods is shown in Figs. 3–5.

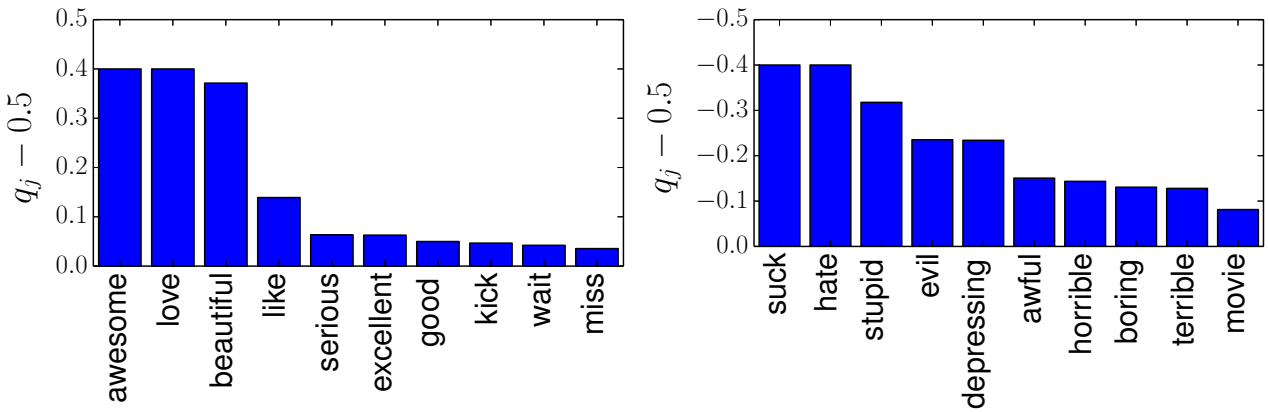


Figure 3 $q_j - 0.5$ for most relevant tags according to RTM-MAP-EM

The RVM chose a lot of rare words to represent the negative class (words "crappy," "shitty," "lousy," "blame," "afraid," "piece," and "idiot" have less than seven occurrences in the dataset) and the words from the positive class does not look relevant at all. The RVM failed to solve the relevance determination problem on this dataset.

The words chosen by RTM-MAP-EM and logistic regression are quite intuitive in case of this problem. the present model, it isn't true. Most of them are very emotional. Top-20 words chosen by logistic regression are almost the same as top-20 words, chosen by the present method. However, the present model provided a more sparse solution with comparable prediction performance. Therefore, the present method proved to be better at relevance determination than logistic regression and the RVM on this dataset.

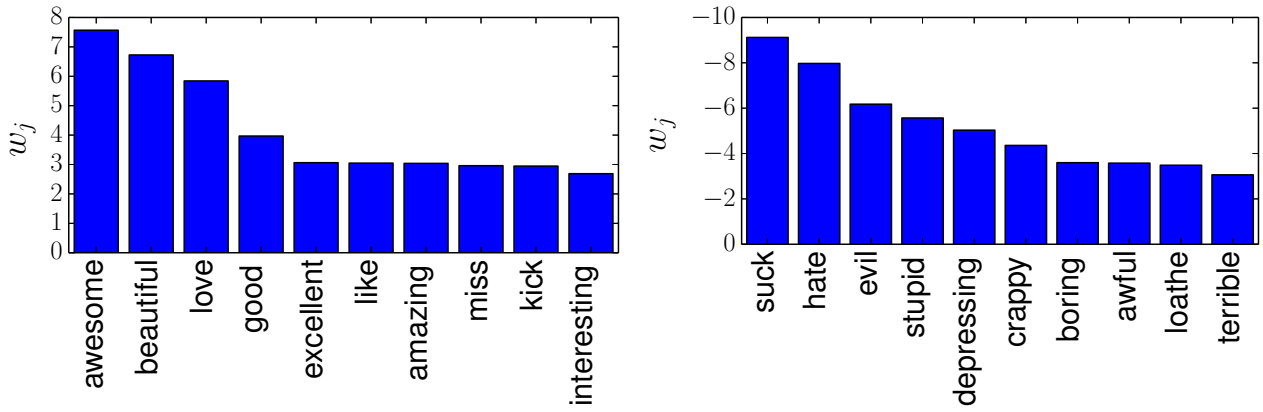


Figure 4 Weights of linear model tuned by L1-LR

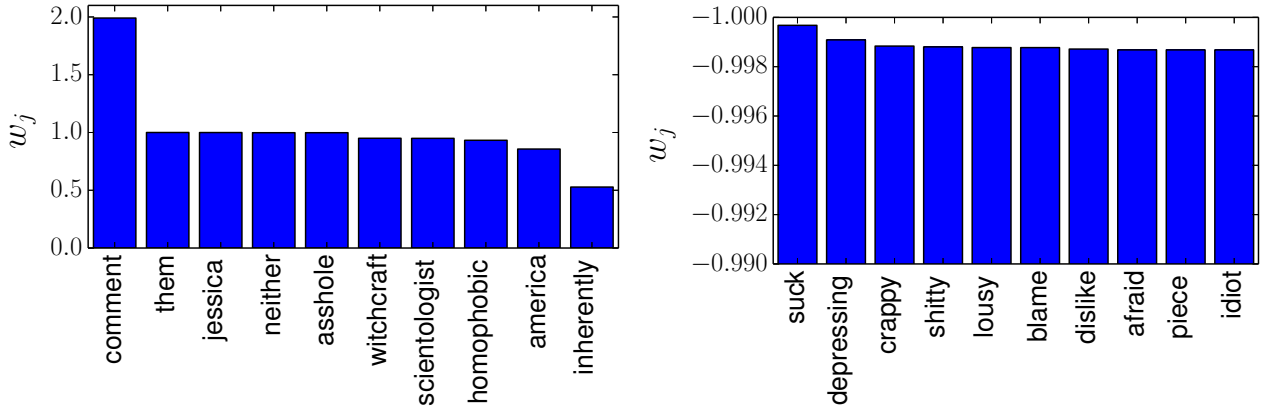


Figure 5 Weights of linear model tuned by RVM

5 Concluding Remarks

Most of previous work on Bayesian ARD approach consider only Gaussian prior. The authors demonstrate that other appropriate priors may also work well. It means that Bayesian ARD approach might be more broad than it was considered before and is not limited to the usage of Gaussian prior. Also, a method to solve a binary classification problem with binary features is suggested and an experimental comparison which shows that the present model is comparable to the state-of-the-art methods of classification and feature selection is provided. The experiments show that the present model provides better feature selection results than the classic feature selection models like RVM and L1-LR.

Appendix A

Proof of Theorem 1

Derive a variational lower bound on the likelihood function for a single object $\boldsymbol{x}$. Let Q be the set of its tags: $Q = \{j|x_j = 1\}$. After some transformations and a change of variables, a convex function is obtained and its tangent is used as its variational lower bound:

$$\mathsf{P}(t \,|\, \boldsymbol{x}, \boldsymbol{q}) = \frac{\prod\limits_{j\in Q} q_j^t (1-q_j)^{1-t}}{\prod\limits_{j\in Q} q_j + \prod\limits_{j\in Q} (1-q_j)} = \left(1 + \prod_{j\in Q} \left(\frac{1-q_j}{q_j}\right)^{2t-1}\right)^{-1}$$

as $t \in \{0, 1\}$;

$$s_j := \left(\frac{1-q_j}{q_j}\right)^{|Q|(2t-1)};$$

so,

$$\log \mathsf{P}(t \,|\, \boldsymbol{x}, \boldsymbol{q}) = -\log\left(1 + \left(\prod_{j\in Q} s_j\right)^{1/|Q|}\right). \tag{6}$$

As $s_j > 0$, the geometric mean $\left(\prod\limits_{j\in Q} s_j\right)^{1/|Q|}$ is concave with respect to $\boldsymbol{s}$ [7, p. 74]. As $f(x) = -\log x$ is convex and nonincreasing, the whole expression on the right part of (6) is convex with respect to $\boldsymbol{s}$ [7, p. 84]. Therefore, its tangent is its variational lower bound and after making inverse change of variables and taking the exponent, one obtains a variational lower bound on $\mathsf{P}(t \,|\, \boldsymbol{x}, \boldsymbol{q})$.

The variational lower bound on the likelihood of an object $\boldsymbol{x}_i$ from the training set looks as follows:

$$\mathsf{P}(t_i \,|\, \boldsymbol{x}_i, \boldsymbol{q}) \geqslant L_i(\boldsymbol{q}, \boldsymbol{\eta}_i) = c_i(\boldsymbol{\eta}_i) \exp\left(\sum_{j\in Q_i} \tilde{c}_{ij}(\boldsymbol{\eta}_i) \left(\frac{1-q_j}{q_j}\right)^{|Q_i|(2t_i-1)}\right)$$

where

$$Q_i = \{j|x_{ij} = 1\};$$

$$c_i(\boldsymbol{\eta}_i) = \frac{\prod\limits_{j\in Q_i} \eta_{ij}^{t_i}(1-\eta_{ij})^{1-t_i}}{\prod\limits_{j\in Q_i} \eta_{ij} + \prod\limits_{j\in Q_i} (1-\eta_{ij})} \exp\left(\frac{\prod\limits_{j\in Q_i} \eta_{ij}^{1-t_i}(1-\eta_{ij})^{t_i}}{\prod\limits_{j\in Q_i} \eta_{ij} + \prod\limits_{j\in Q_i} (1-\eta_{ij})}\right);$$

$$\tilde{c}_{ij}(\boldsymbol{\eta}_i) = -\frac{\prod\limits_{j\in Q_i} \eta_{ij}^{t_i}(1-\eta_{ij})^{1-t_i}}{\prod\limits_{j\in Q_i} \eta_{ij} + \prod\limits_{j\in Q_i} (1-\eta_{ij})} \left(\frac{\eta_{ij}}{1-\eta_{ij}}\right)^{|Q_i|(2t_i-1)} |Q_i|^{-1}.$$

The following equation concludes the proof:

$$E(\boldsymbol{\alpha}) = \int \mathsf{P}(\boldsymbol{t} \,|\, \boldsymbol{X}, \boldsymbol{q}) p(\boldsymbol{q} \,|\, \boldsymbol{\alpha})\, d\boldsymbol{q} \geqslant \int \prod_{i=1}^{n} L_i(\boldsymbol{q}, \boldsymbol{\eta}_i) \prod_{j=1}^{n} p(q_j \,|\, \alpha)\, d\boldsymbol{q} = \tilde{E}(\boldsymbol{\alpha}, \boldsymbol{H}).$$

Note that each object has its own set of variational parameters. As $\boldsymbol{\eta}_{ij}$ is dummy if $q_j \notin Q_i$, there are $\sum_{i,j} x_{ij}$ essential variational parameters.

Appendix B

RTM-MAP-EM algorithm

Algorithm 1 RTM-MAP-EM

Require: training set $(\boldsymbol{X}, \boldsymbol{t})$; maximum number of iterations T; tolerance ε
Ensure: tuned vector of hyperparameters $\boldsymbol{\alpha}$

$\boldsymbol{\alpha}^0 \leftarrow (1, \ldots, 1)^{\mathrm{T}}$ // *Initial value of hyperparameters*
$\boldsymbol{\eta}^0 \leftarrow \arg\max_{\boldsymbol{q}} \sum_{i=1}^{n} \log \mathsf{P}(t_i \,|\, \boldsymbol{x}_i, \boldsymbol{q}) + \log \mathsf{P}(\boldsymbol{q} \,|\, \boldsymbol{\alpha}^0)$ // $\boldsymbol{\eta} = \boldsymbol{q}^{\mathrm{MAP}}$
for $k = 0$ to T
  // *E-step:*
  $\boldsymbol{\eta}^k \leftarrow \arg\max_{\boldsymbol{q}} \sum_{i=1}^{n} \log \mathsf{P}(t_i \,|\, \boldsymbol{x}_i, \boldsymbol{q}) + \log \mathsf{P}(\boldsymbol{q} \,|\, \boldsymbol{\alpha}^{k-1})$ // $\boldsymbol{\eta} = \boldsymbol{q}^{\mathrm{MAP}}$
  $\boldsymbol{H}^k \leftarrow (\boldsymbol{\eta}^k, \ldots, \boldsymbol{\eta}^k)^T$
  // *M-step:*
  $\boldsymbol{\alpha}^k \leftarrow \arg\max_{\boldsymbol{\alpha}} \log \tilde{E}(\boldsymbol{\alpha}, \boldsymbol{H}^k)$
  if $\|\boldsymbol{\alpha}^k - \boldsymbol{\alpha}^{k-1}\|^2 < \varepsilon$ **then**
    break
$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha}^k$
return $\boldsymbol{\alpha}$

Some notes about implementation:

- q_j and η_{ij} are bound to $[0.1, 0.9]$ for all i, j. Otherwise, computations tend to be unstable;
- $k_E = k_M = 10$ in all experiments;
- α_j is bound to $[0, 1000]$ for all j. If $\alpha_j \geqslant 900$, it is considered to be infinite: $\alpha_j := +\infty$; and
- there seems to be no need to wait till full convergence; for RTM-MAP-EM, $T = 20$ was enough in both experiments.

References

[1] MacKay, D., and R. Neal. 1994. Automatic relevance determination for neural networks. Cambridge University. Technical Report.

[2] Tipping, M. E. 2001. Sparse Bayesian learning and the relevance vector machine. *J. Machine Learning Res.* 1:211–244.

[3] MacKay, D. 1992. Bayesian interpolation. *Neural Computation* 4:415–447.

[4] Bishop, C. M. 2006. *Pattern recognition and machine learning*. New York, NY: Springer. 738 p.

[5] Salakhutdinov, R., S. Roweis, and Z. Ghahramani. 2002. On the convergence of bound optimization algorithms. *19th Conference on Uncertainty in Artificial Intelligence Proceedings*. 10:509–516.

[6] Jaakkola, T. S., and M. I. Jordan. 2000. Bayesian logistic regression: A variational approach. *Stat. Comput.* 10:25–37.

[7] Boyd, S., and L. Vandenberghe. 2004. *Convex optimization*. Cambridge: Cambridge University Press. 716 p.

[8] Byrd, R., P. Lu, and J. Nocedal. 1995. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Stat. Comput.* 16(5):1190–1208.

[9] Kaggle in Class. 2011. UMICH SI650 — Sentiment Classification. Available at: `http://inclass.kaggle.com/c/si650winter11/data` (accessed December 29, 2015).

Received June 14, 2015

Литература

[1] *MacKay D., Neal R.* Automatic relevance determination for neural networks // Cambridge University, 1994. Technical Report.

[2] *Tipping M. E.* Sparse Bayesian learning and the relevance vector machine // J. Machine Learning Res., 2001. No. 1. P. 211–244.

[3] *MacKay D.* Bayesian interpolation // Neural Computation, 1992. No 4. P. 415–447.

[4] *Bishop C. M.* Pattern recognition and machine learning. — New York, NY, USA: Springer, 2006. 738 p.

[5] *Salakhutdinov R., Roweis S., Ghahramani Z.* On the convergence of bound optimization algorithms // 19th Conference on Uncertainty in Artificial Intelligence Proceedings, 2002. No. 10. P. 509–516.

[6] *Jaakkola T. S., Jordan M. I.* Bayesian logistic regression: A variational approach // Stat. Comput., 2000. No. 10. P. 25–37.

[7] *Boyd S., Vandenberghe L.* Convex optimization. — Cambridge: Cambridge University Press, 2004. 716 p.

[8] *Byrd R., Lu P., Nocedal J.* A limited memory algorithm for bound constrained optimization // SIAM J. Sci. Stat. Comput., 1995. Vol. 16, No. 5. P. 1190–1208.

[9] Kaggle in Class. UMICH SI650 — Sentiment Classification. 2011. `http://inclass.kaggle.com/c/si650winter11/data`.

Поступила в редакцию 14.06.2015