POLITECHNIKA KOSZALIŃSKA

Zeszyty Naukowe
Wydziału Elektroniki i Informatyki

Nr 5

KOSZALIN 2013

Zeszyty Naukowe Wydziału Elektroniki i Informatyki Nr 5

ISSN 1897-7421

Projekt okładki
Tadeusz Walczak

Skład, łamanie
Maciej Bączek

Wydawnictwo Uczelniane Politechniki Koszalińskiej
75-620 Koszalin, ul. Racławicka 15-17

Koszalin 2013, wyd. I, ark. wyd. 7,1, format B-5, nakład 100 egz.
Druk: INTRO-DRUK, Koszalin

Spis treści

Dariusz Jakóbczak
Katedra Podstaw Informatyki i Zarządzania
Wydział Elektroniki i Informatyki
Politechnika Koszalińska

Active Object Modeling and Applications via the Method of Hurwitz-Radon Matrices

Keywords: Hurwitz-Radon Matrices, Active Object Modeling

1. Introduction

Object description by its contour is a critical part in many applications of image processing. Thus computer vision and artificial intelligence have a problem: how to model the active shape [1,2] via discrete set of two-dimensional boundary points? Also subject of shape representation and shape description is still opened [3,4]. The author wants to approach a problem of image structure representation by characteristic contour points and not limited to closed curves, but also working on open curves (for example a signature or handwriting). Proposed method relies on active functional modeling of boundary points situated between the basic set of the nodes. The functions that are used in calculations represent whole family of elementary functions: trigonometric, cyclometric, logarithmic, exponential and power function. Nowadays methods apply mainly polynomial functions, for example Bernstein polynomials in Bezier curves, splines and NURBS [5]. Numerical methods for data interpolation are based on polynomial or trigonometric functions, for example Lagrange, Newton, Aitken and Hermite methods. These methods have some weak sides [6] and are not sufficient for object modeling in the situations when the shape cannot be build by polynomials or trigonometric functions. Also trigonometric basis functions in Fourier Series Shape Models are not appropriate for describing all shapes. Model-based vision such as Active Contour Models (called Snakes) or Active Shape Models use the training sets to fit the data and they are applied only for closed curves. In this paper discussed approach is not limited to closed curves and it does not use a training set of some images, but only a set of two-dimensional nodes of the curve. Proposed Active Object Modeling is the functional modeling via any elementary functions and it helps us to fit the contour and to match the shape in object modeling or image analysis. The author presents novel method of flexible modeling and building the image structure for applications in signature and handwriting modeling, curve fitting, object representation and shape geometry.

This paper takes up new method of two-dimensional Active Object Modeling (AOM) by using a family of Hurwitz-Radon matrices. The method of Hurwitz-Radon Matrices (MHR) requires minimal assumptions about object. The only information about shape or curve is the set of at least five nodes. Proposed method of Hurwitz-Radon Matrices (MHR) is applied in curve modeling via different coefficients: sinusoidal, cosinusoidal, tangent, logarithmic, exponential, arc sin, arc cos, arc tan or power. Function for coefficient calculations is chosen individually at each Active Object Modeling and it depends on initial requirements and shape specifications to fit and to match the object. MHR method uses two-dimensional vectors (x,y) for data analysis and curve modeling. Shape of the object is represented by succeeding boundary points $(x_i,y_i) \in \boldsymbol{R}^2$ as follows in MHR method:

1. At least five nodes (x_1,y_1), (x_2,y_2), (x_3,y_3), (x_4,y_4) and (x_5,y_5) if MHR method is implemented with matrices of dimension $N = 2$;
2. For better modeling nodes ought to be settled at key points of the curve, for example local minimum or maximum and at least one point between two successive local extrema.

Condition 1 is connected with important features of MHR method: MHR version with matrices of dimension $N = 2$ (MHR-2) needs at least five nodes, MHR version with matrices of dimension $N = 4$ (MHR-4) needs at least nine nodes and MHR version with matrices of dimension $N = 8$ (MHR-8) needs at least 17 nodes. Condition 2 means for example the highest point of the object in a particular orientation, convexity changing or curvature extrema. So this paper wants to answer the question: how to model the active object for discrete set of points?

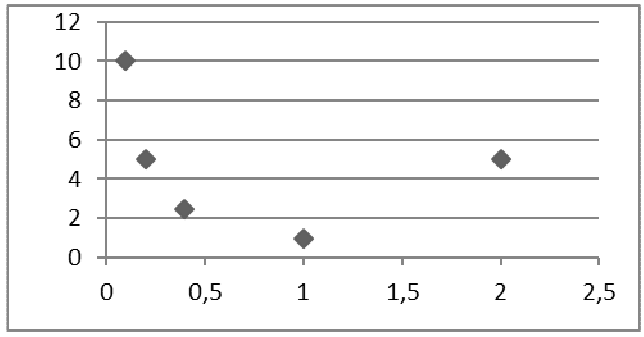


Fig. 1. Nodes of the object before modelin

Coefficients for Active Object Modeling are computed via individual features of the object boundary using power function, sinus, cosine, tangent, logarithm, exponent or arc sin, arc cos, arc tan.

2. Active Object Modeling via MHR

The method of Hurwitz – Radon Matrices (MHR), described in this paper, is computing points between two successive nodes of the curve. Data of Active Object Modeling are interpolated and parameterized for real number $\alpha \in [0;1]$ in the range of two successive nodes. MHR calculations are introduced with square matrices of dimension $N = 2$, 4 or 8. Matrices A_i, $i = 1,2\ldots m$ satisfying

$A_jA_k+A_kA_j = 0$, $A_j^2 = -I$ for $j \neq k$; $j, k = 1,2...m$

are called *a family of Hurwitz - Radon matrices*, discussed by Adolf Hurwitz and Johann Radon separately in 1923. A family of Hurwitz - Radon (HR) matrices [7] are skew-symmetric ($A_i^{\mathrm{T}} = -A_i$), $A_i^{-1} = -A_i$ and only for dimension N=2, 4 or 8 the family of HR matrices consists of N-1 matrices. For $N = 2$:

$$A_1 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

For $N = 4$ there are three HR matrices with integer entries:

$$A_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}.$$

For $N = 8$ we have seven HR matrices with elements 0, ±1. So far HR matrices have found applications in Space-Time Block Coding (STBC) [8] and orthogonal design [9], in signal processing [10] and Hamiltonian Neural Nets [11].

How coordinates of curve points are applied in Active Object Modeling? If boundary points have the set of following nodes $\{(x_i,y_i)$, $i = 1, 2, \ldots, n\}$ then HR matrices combined with the identity matrix I_N are used to build the orthogonal Hurwitz - Radon Operator (OHR). For points $p_1=(x_1,y_1)$ and $p_2=(x_2,y_2)$ OHR of dimension $N = 2$ is build via matrix M_2:

$$M_2(p_1, p_2) = \frac{1}{x_1^2 + x_2^2} \begin{bmatrix} x_1y_1 + x_2y_2 & x_2y_1 - x_1y_2 \\ x_1y_2 - x_2y_1 & x_1y_1 + x_2y_2 \end{bmatrix} \tag{1}$$

For points $p_1=(x_1,y_1)$, $p_2=(x_2,y_2)$, $p_3=(x_3,y_3)$ and $p_4=(x_4,y_4)$ OHR M_4 of dimension $N = 4$ is introduced:

$$M_4(p_1, p_2, p_3, p_4) = \frac{1}{x_1^2 + x_2^2 + x_3^2 + x_4^2} \begin{bmatrix} u_0 & u_1 & u_2 & u_3 \\ -u_1 & u_0 & -u_3 & u_2 \\ -u_2 & u_3 & u_0 & -u_1 \\ -u_3 & -u_2 & u_1 & u_0 \end{bmatrix} \tag{2}$$

where

$$u_0 = x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4\,, \quad u_1 = -x_1y_2 + x_2y_1 + x_3y_4 - x_4y_3\,,$$
$$u_2 = -x_1y_3 - x_2y_4 + x_3y_1 + x_4y_2\,, \quad u_3 = -x_1y_4 + x_2y_3 - x_3y_2 + x_4y_1\,.$$

For nodes $p_1=(x_1,y_1)$, $p_2=(x_2,y_2)$,… and $p_8=(x_8,y_8)$ OHR M_8 of dimension $N = 8$ is constructed [12] similarly as (1) and (2):

$$M_8(p_1, p_2 \ldots p_8) = \frac{1}{\sum_{i=1}^{8} x_i^2} \begin{bmatrix} u_0 & u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 \\ -u_1 & u_0 & u_3 & -u_2 & u_5 & -u_4 & -u_7 & u_6 \\ -u_2 & -u_3 & u_0 & u_1 & u_6 & u_7 & -u_4 & -u_5 \\ -u_3 & u_2 & -u_1 & u_0 & u_7 & -u_6 & u_5 & -u_4 \\ -u_4 & -u_5 & -u_6 & -u_7 & u_0 & u_1 & u_2 & u_3 \\ -u_5 & u_4 & -u_7 & u_6 & -u_1 & u_0 & -u_3 & u_2 \\ -u_6 & u_7 & u_4 & -u_5 & -u_2 & u_3 & u_0 & -u_1 \\ -u_7 & -u_6 & u_5 & u_4 & -u_3 & -u_2 & u_1 & u_0 \end{bmatrix} \tag{3}$$

where

$$\underline{u} = \begin{bmatrix} y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 & y_8 \\ -y_2 & y_1 & -y_4 & y_3 & -y_6 & y_5 & y_8 & -y_7 \\ -y_3 & y_4 & y_1 & -y_2 & -y_7 & -y_8 & y_5 & y_6 \\ -y_4 & -y_3 & y_2 & y_1 & -y_8 & y_7 & -y_6 & y_5 \\ -y_5 & y_6 & y_7 & y_8 & y_1 & -y_2 & -y_3 & -y_4 \\ -y_6 & -y_5 & y_8 & -y_7 & y_2 & y_1 & y_4 & -y_3 \\ -y_7 & -y_8 & -y_5 & y_6 & y_3 & -y_4 & y_1 & y_2 \\ -y_8 & y_7 & -y_6 & -y_5 & y_4 & y_3 & -y_2 & y_1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix} \tag{4}$$

and $\underline{u} = (u_0, u_1,\ldots, u_7)^T$ (4). OHR operators M_N (1)-(3) satisfy the condition of interpolation

$$M_N \cdot \mathbf{x} = \mathbf{y} \tag{5}$$

for $\mathbf{x} = (x_1,x_2\ldots,x_N)^T \in \boldsymbol{R}^N$, $\mathbf{x} \neq \mathbf{0}$, $\mathbf{y} = (y_1,y_2\ldots,y_N)^T \in \boldsymbol{R}^N$ and N =2, 4 or 8.

2.1 Functional coefficients in MHR Active Object Modeling

Coordinates of points settled between the nodes are computed [13] using described MHR method [14]. Each real number $c \in [a;b]$ is calculated by a convex combination $c = \alpha \cdot a + (1 - \alpha) \cdot b$ for

$$\alpha = \frac{b-c}{b-a} \in [0;1]. \tag{6}$$

The average OHR operator M of dimension $N = 2$, 4 or 8 is build:

$$M = \gamma \cdot A + (1-\gamma) \cdot B . \tag{7}$$

The OHR matrix A is constructed (1)-(3) by every second point $p_1=(x_1=a,y_1)$, $p_3=(x_3,y_3)$,… and $p_{2N-1}=(x_{2N-1},y_{2N-1})$:

$$A = M_N (p_1, p_{3,\ldots,} p_{2N-1}).$$

The OHR matrix B is computed (1)-(3) by data $p_2=(x_2=b,y_2)$, $p_4=(x_4,y_4)$,… and $p_{2N}=(x_{2N},y_{2N})$:

$$B = M_N(p_2, p_{4,\ldots,} p_{2N}).$$

Vector of first coordinates C is defined for

$$c_i = \alpha \cdot x_{2i-1} + (1-\alpha) \cdot x_{2i} \quad , \quad i = 1, 2, \ldots, N \tag{8}$$

and $C = [c_1, c_2, \ldots, c_N]^T$. The formula to calculate second coordinates $y(c_i)$ is similar to the interpolation formula (5):

$$Y(C) = M \cdot C \tag{9}$$

where $Y(C) = [y(c_1), y(c_2), \ldots, y(c_N)]^T$. So modeled value of $y(c_i)$ depends on four, eight or sixteen ($2N$) successive nodes, not only two.

Key question is dealing with coefficient γ in (7). Coefficient γ is calculated using different functions (power, sinus, cosine, tangent, logarithm, exponent, arc sin, arc cos, arc tan) and choice of function is connected with initial requirements and shape specifications during fitting and matching of the object. Coefficients γ and α (6) are strongly related:

1. $\gamma = 0 \leftrightarrow \alpha = 0$;
2. $\gamma = 1 \leftrightarrow \alpha = 1$;
3. $\gamma \in [0;1]$.

Different values of coefficient γ are connected with implemented functions and positive real number s:

$\gamma=\alpha^s$, $\gamma=sin(\alpha^s \cdot \pi/2)$, $\gamma=sin^s(\alpha \cdot \pi/2)$, $\gamma=1-cos(\alpha^s \cdot \pi/2)$, $\gamma=1-cos^s(\alpha \cdot \pi/2)$, $\gamma=tan(\alpha^s \cdot \pi/4)$, $\gamma=tan^s(\alpha \cdot \pi/4)$, $\gamma=log_2(\alpha^s+1)$, $\gamma=log_2^s(\alpha+1)$, $\gamma=(2^\alpha-1)^s$, $\gamma=2/\pi \cdot arcsin(\alpha^s)$, $\gamma=(2/\pi \cdot arcsin\alpha)^s$, $\gamma=1-2/\pi \cdot arccos(\alpha^s)$, $\gamma=1-(2/\pi \cdot arccos\alpha)^s$, $\gamma=4/\pi \cdot arctan(\alpha^s)$, $\gamma=(4/\pi \cdot arctan\alpha)^s$, $\gamma=ctg(\pi/2-\alpha^s \cdot \pi/4)$, $\gamma=ctg^s(\pi/2-\alpha \cdot \pi/4)$, $\gamma=2-4/\pi \cdot arcctg(\alpha^s)$, $\gamma=(2-4/\pi \cdot arcctg\alpha)^s$.

For example if $s = 1$ then:

basic MHR $\gamma = \alpha$, $\gamma = sin(\alpha \cdot \pi/2)$, $\gamma = 1-cos(\alpha \cdot \pi/2)$, $\gamma = tan(\alpha \cdot \pi/4)$, $\gamma = log_2(\alpha + 1)$, $\gamma = 2^\alpha - 1$, $\gamma = 2/\pi \cdot arcsin(\alpha)$, $\gamma = 1-2/\pi \cdot arccos(\alpha)$, $\gamma = 4/\pi \cdot arctan(\alpha)$.

What is very important, above functions used in γ calculations are strictly monotonic for $\alpha \in [0;1]$, because $\gamma \in [0;1]$ too. Choice of function and parameter s depends on object specifications and individual requirements. Fixing of unknown coordinates for curve points using (6)-(9) is called by author the method of Hurwitz - Radon Matrices (MHR) [15]. Each strictly monotonic function between points (0;0) and (1;1) can be used in Active Object Modeling – not only above functions.

3. Applications of AOM in Handwriting Modeling

Boundary nodes: (0.1;10), (0.2;5), (0.4;2.5), (1;1) and (2;5) from Fig.1 are used in some examples of MHR Active Object Modeling with different γ. These examples are connected with handwriting modeling for letter "*w*". Points of the object are calculated with matrices of dimension $N = 2$ and $\alpha = 0.1, 0.2, \ldots, 0.9$.

Example 1

Sinusoidal modeling with $\gamma = sin(\alpha \cdot \pi/2)$.

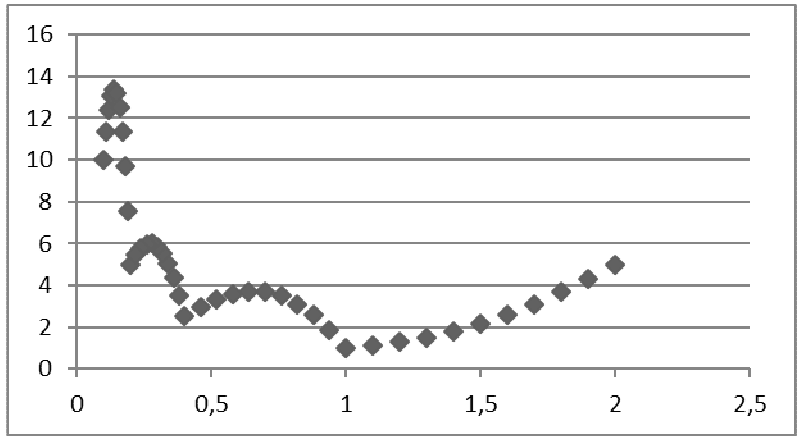


Fig. 2. Sinusoidal modeling with nine reconstructed points between nodes

Example 2

Tangent modeling for $\gamma = tan(\alpha \cdot \pi/4)$.

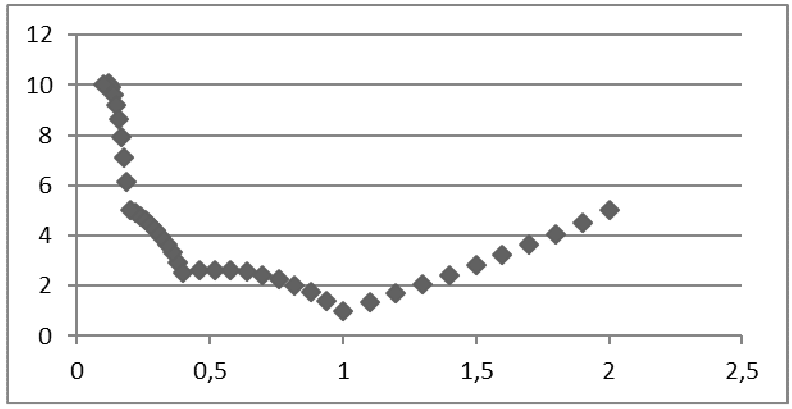


Fig. 3. Tangent modeling with nine interpolated boundary points between nodes

Example 3

Tangent modeling with $\gamma = tan(\alpha^s \cdot \pi/4)$ and $s = 1.5$.

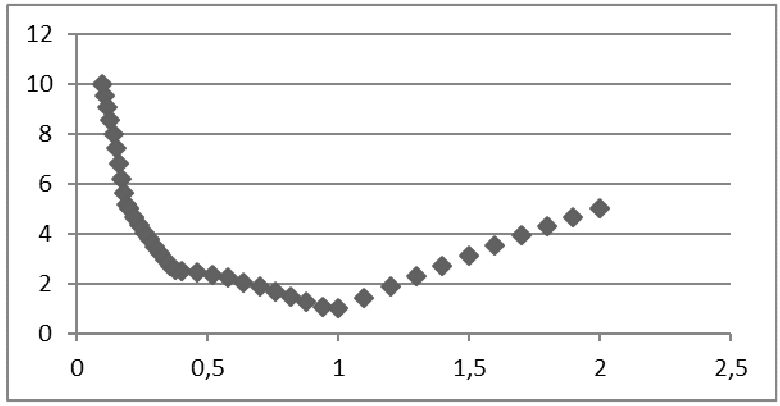


Fig. 4. Tangent modeling with nine recovered curve points between nodes

Example 4

Tangent modeling for $\gamma = tan(\alpha^s \cdot \pi/4)$ and $s = 1.797$.

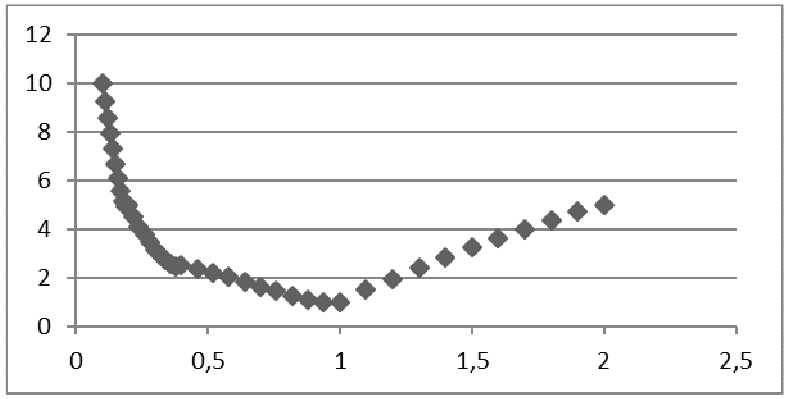


Fig. 5. Tangent modeling with nine reconstructed points between nodes

Example 5

Sinusoidal modeling with $\gamma = sin(\alpha^s \cdot \pi/2)$ and $s = 2.759$.

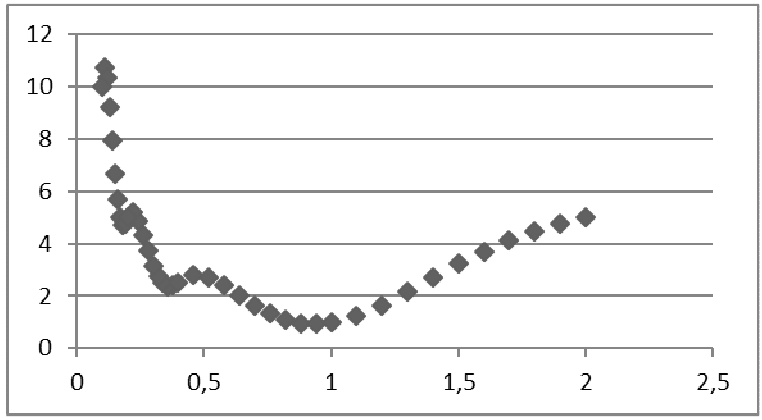


Fig. 6. Sinusoidal modeling with nine interpolated curve points between nodes

Example 6

Power function modeling for $\gamma = \alpha^s$ and $s = 2.1205$.

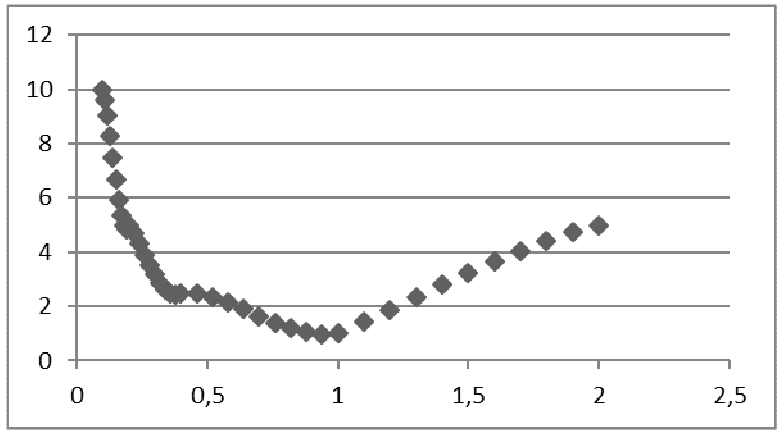


Fig. 7. Power function modeling with nine recovered object points between nodes

Example 7

Logarithmic modeling with $\gamma = log_2(\alpha^s + 1)$ and $s = 2.533$.

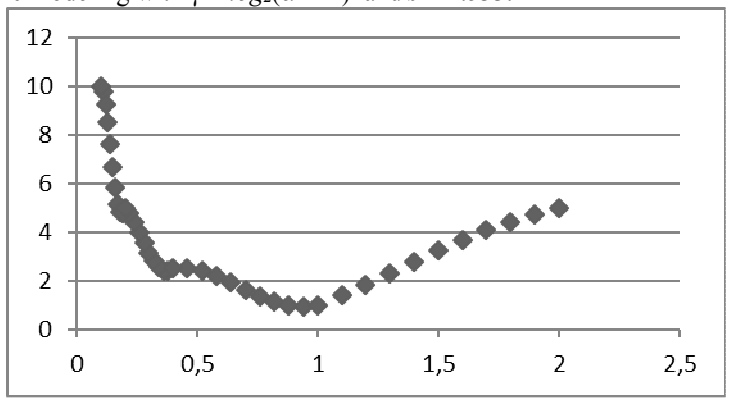


Fig. 8. Logarithmic modeling with nine reconstructed points between nodes

These seven examples demonstrate possibilities of Active Object Modeling for boundary nodes. Reconstructed values and interpolated points, calculated by MHR method, are applied in the process of curve modeling for fitting and matching the object during its analysis. Recovered points can be treated as a part of signature or handwriting and used during different stages of image processing, for example signature modeling, object representation, shape geometry and curve fitting. Every individual signature or handwriting, each letter or number can be modeled by some function for parameter γ. This parameter is treated as characteristic feature of letter or figure.

4. Conclusions

The method of Hurwitz-Radon Matrices (MHR) enables active modeling of two-dimensional shapes using different coefficients γ: sinusoidal, cosinusoidal, tangent, logarithmic, exponential, arc sin, arc cos, arc tan or power function [16]. Function for γ calculations is chosen individually at each Active Object Modeling (AOM) and depends on initial requirements and curve specifications. MHR method leads to shape modeling via discrete set of fixed points. So MHR makes possible the combination of two important problems: interpolation and modeling. Main features of MHR method are:

a) modeling of L points is connected with the computational cost of rank $O(L)$;
b) MHR is well-conditioned method (orthogonal matrices) [17];
c) coefficient γ is crucial in the process of AOM and it is computed individually for each object (contour, letter or figure).

Future works are going to: features of coefficient γ, implementation of MHR and AOM in object recognition [18], shape representation, curve fitting, contour modeling and parameterization [19].

References

1. Mahoor, M.H., Abdel-Mottaleb, M., Nasser Ansari, A.: Improved Active Shape Model for Facial Feature Extraction in Color Images. Journal of Multimedia 1(4), 21-28 (2006)
2. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active Shape Models-Their Training and Application. Computer Vision and Image Understanding 1(61), 38-59 (1995)
3. Amanatiadis, A., Kaburlasos, V.G., Gasteratos, A., Papadakis, S.E.: Evaluation of Shape Descriptors for Shape-Based Image Retrieval. IET Image Processing 5(5), 493-499 (2011)
4. Zhang, D., Lu, G.: Review of Shape Representation and Description Techniques. Pattern Recognition 1(37), 1-19 (2004)
5. Schumaker, L.L.: Spline Functions: Basic Theory. Cambridge Mathematical Library (2007)
6. Dahlquist, G., Bjoerck, A.: Numerical Methods. Prentice Hall, New York (1974)
7. Eckmann, B.: Topology, Algebra, Analysis- Relations and Missing Links. Notices of the American Mathematical Society 5(46), 520-527 (1999)
8. Citko, W., Jakóbczak, D., Sieńko, W.: On Hurwitz - Radon Matrices Based Signal Processing. Workshop Signal Processing at Poznan University of Technology (2005)
9. Tarokh, V., Jafarkhani, H., Calderbank, R.: Space-Time Block Codes from Orthogonal Designs. IEEE Transactions on Information Theory 5(45), 1456-1467 (1999)
10. Sieńko, W., Citko, W., Wilamowski, B.: Hamiltonian Neural Nets as a Universal Signal Processor. 28th Annual Conference of the IEEE Industrial Electronics Society IECON (2002)
11. Sieńko, W., Citko, W.: Hamiltonian Neural Net Based Signal Processing. The International Conference on Signal and Electronic System ICSES (2002)
12. Jakóbczak, D.: 2D and 3D Image Modeling Using Hurwitz-Radon Matrices. Polish Journal of Environmental Studies 4A(16), 104-107 (2007)
13. Jakóbczak, D.: Shape Representation and Shape Coefficients via Method of Hurwitz-Radon Matrices. Lecture Notes in Computer Science 6374 (Computer Vision and Graphics: Proc. ICCVG 2010, Part I), Springer-Verlag Berlin Heidelberg, 411-419 (2010)
14. Jakóbczak, D.: Curve Interpolation Using Hurwitz-Radon Matrices. Polish Journal of Environmental Studies 3B(18), 126-130 (2009)
15. Jakóbczak, D.: Application of Hurwitz-Radon Matrices in Shape Representation. In: Banaszak, Z., Świć, A. (eds.) Applied Computer Science: Modelling of

Production Processes 1(6), pp. 63-74. Lublin University of Technology Press, Lublin (2010)

16. Jakóbczak, D.: Object Modeling Using Method of Hurwitz-Radon Matrices of Rank k. In: Wolski, W., Borawski, M. (eds.) Computer Graphics: Selected Issues, pp. 79-90. University of Szczecin Press, Szczecin (2010)
17. Jakóbczak, D.: Implementation of Hurwitz-Radon Matrices in Shape Representation. In: Choraś, R.S. (ed.) Advances in Intelligent and Soft Computing 84, Image Processing and Communications: Challenges 2, pp. 39-50. Springer-Verlag, Berlin Heidelberg (2010)
18. Jakóbczak, D.: Object Recognition via Contour Points Reconstruction Using Hurwitz-Radon Matrices. In: Józefczyk, J., Orski, D. (eds.) Knowledge-Based Intelligent System Advancements: Systemic and Cybernetic Approaches, pp. 87-107. IGI Global, Hershey PA, USA (2011)
19. Jakóbczak, D.: Curve Parameterization and Curvature via Method of Hurwitz-Radon Matrices. Image Processing & Communications- An International Journal 1-2(16), 49-56 (2011)

Abstract

Artificial intelligence and computer vision need methods for object modeling having discrete set of boundary points. A novel method of Hurwitz-Radon Matrices (MHR) is used in shape modeling. Proposed method is based on the family of Hurwitz-Radon (HR) matrices which possess columns composed of orthogonal vectors. Two-dimensional active curve is modeling via different functions: sinus, cosine, tangent, logarithm, exponent, arc sin, arc cos, arc tan and power function. It is shown how to build the orthogonal matrix OHR operator and how to use it in a process of object modeling.

Streszczenie

Matematyka i jej zastosowania wymagają odpowiednich metod modelowania oraz interpolacji danych. Autorska metoda Macierzy Hurwitza-Radona (MHR) jest sposobem modelowania krzywej 2D. Oparta jest ona na rodzinie macierzy Hurwitza-Radona, których kluczową cechą jest ortogonalność kolumn. Dwuwymiarowe dane są interpolowane z wykorzystaniem różnych funkcji rozkładu prawdopodobieństwa: potęgowych, wielomianowych, wykładniczych, logarytmicznych, trygonometrycznych, cyklometrycznych. W pracy pokazano budowę ortogonalnego operatora macierzowego i jego wykorzystanie w rekonstrukcji i modelowaniu danych.

Słowa kluczowe: macierze Hurwitza-Radona, aktywne modelowanie obiektów

Mirosław Andrzej Maliński
Łukasz Barłomiej Chrobak
Leszek Bychto
Department of Electronics and Computer Science
Technical University of Koszalin
2 Śniadeckich St.
75-343 Koszalin
Poland

Michał Pawlak
Faculty of Physics, Astronomy and Informatics
Nicolaus Copernicus University
5 Grudziądzka St.
87-100 Toruń
Poland

Investigations of the Implanted Layer in Silicon Based on the Modulated Free Carrier Absorption Phenomenon

Keywords: semiconductors, ion implantation, nondestructive testing techniques

Introduction

Modulated free carrier absorption technique uses the effect of absorption of the infrared light by free electrons. In this method free electrons are generated by the modulated in intensity beam of light of the energy of photons bigger than the energy gap of the semiconductor. As a result a periodical concentration of electrons is observed in the conduction band. At the same time the same spot of the sample is illuminated by the constant in intensity infrared (IR) beam of light. Periodical changing concentration of electrons modulates transmission of the IR beam of light. These changes of the intensity of the transmitted light are detected by the photodiode. Voltage signal of the photodiode is amplified and measured by the lock-in amplifier. Typically amplitude and phase frequency characteristics of this signal are measured and from the fitting of the theoretical characteristics to the experimental ones the lifetime of the optically generated

carriers is measured. Such a method was applied for example for investigations of the lifetime of carriers in silicon and the results published in papers [2, 3]. Basics of the MFCA method are described in paper [4]. Different modifications of the MFCA method are described in papers [5-8]. The characteristic feature of these papers is that they deal with the measurements of the life time of carriers in silicon and they describe uniform samples e.g. nonimplanted which can be analyzed in the single layer model. The paper is the attempt of the application of the MFCA method for investigations of the implanted silicon samples. In this case the physical model is a two layer one: a thin implanted layer of the thickness d on a substrate of the thickness L. Thickness of the implanted layer is much smaller than the thickness of the sample. In this paper a comparison of the frequency amplitude characteristics of the MFCA signal of the nonimplanted and O^{+6} and Ar^{+8} implanted samples is presented. The aim of the investigations was to check the influence of the implantation on the frequency characteristics of the MFCA signal. In this paper a possible explanation of the observed effect is given in the model of a damaged layer.

Sample Preparation and Experimental Set Up

The investigated samples of silicon were both of the n and p type nonimplanted and implanted. Si – p type samples were nonimplanted and implanted by Ar^{+8} ions with the energy of 120 keV and implantation doses $5 \cdot 10^{13}$ ions/cm^2, $5 \cdot 10^{14}$ ions/cm^2 and $5 \cdot 10^{15}$ ions/cm^2. The thickness of the samples l=280 μm, resistivity ρ=5000 Ωcm.

Si – n type samples were nonimplanted and implanted by O^{+6} ions with the energy 90 keV and implantation doses $5 \cdot 10^{13}$ ions/cm^2 and $5 \cdot 10^{14}$ ions/cm^2. The thickness of the samples l=410 μm, resistivity ρ=3-5 Ωcm.

The experimental set up for investigations of the implanted silicon samples consisted of 150 W halogen lamp as a source of the probe light, 660 nm red laser diode as a source of the pump light, beam splitter, adapter for the sample mounting which was connected through the fiber with a detector, lock-in type amplifier and the computer which controlled the measuring process. The schematic diagram of experimental set up is presented in Fig.1. All measurements were performed in the room temperature.

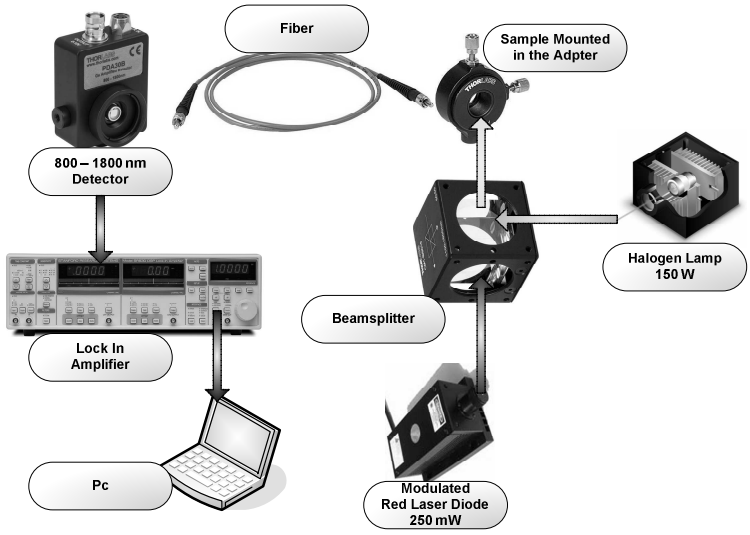


Fig. 1. The experimental set up used for investigations of the implanted layer in silicon samples

Experimental results

Experimental characteristics obtained for silicon samples: nonimplanted and O^{+6} implanted are shown in Fig.2.

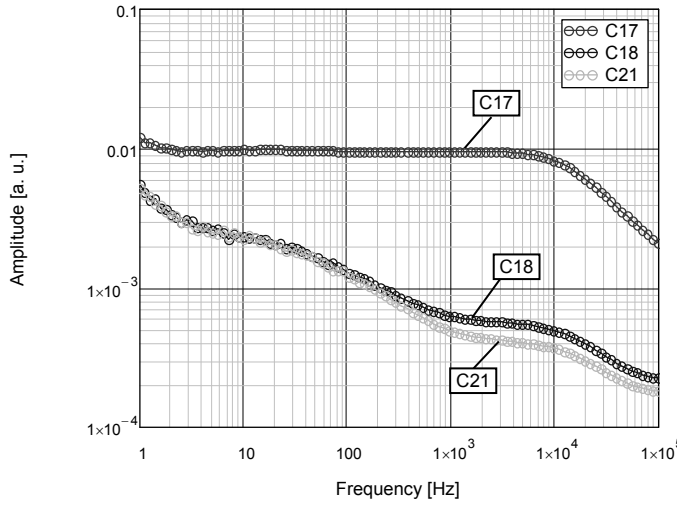


Fig. 2. Experimental MFCA characteristics for n-type silicon samples implanted with O^{+6} ions. C17 - n-type silicon sample nonimplanted, C18 – n-type silicon sample implanted with O^{+6} ions of the energy 90 keV and the dose $5 \cdot 10^{13}$ ions/cm^2, C21 – n-type silicon sample implanted with O^{+6} ions of the energy 90 keV and the dose $5 \cdot 10^{14}$ ions/cm^2.

Experimental characteristics obtained for silicon samples: nonimplanted and Ar^{+8} implanted silicon are shown in Fig.3.

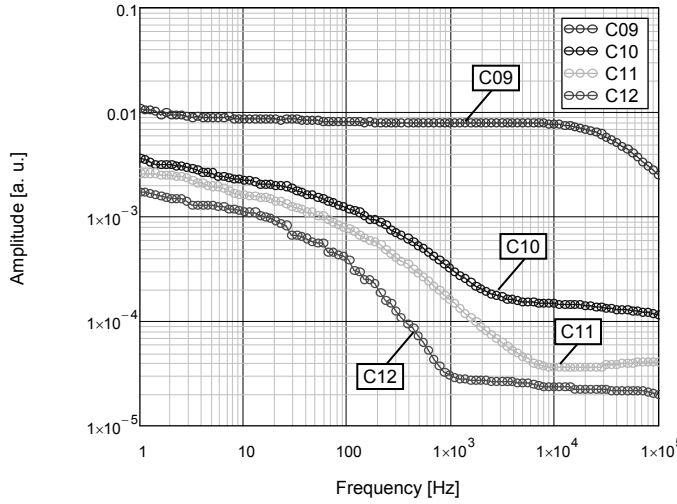


Fig. 3. Experimental MFCA characteristics for p-type silicon samples implanted with Ar^{+8} ions. C09 - p-type silicon sample nonimplanted, C10 – p-type silicon sample implanted with Ar^{+8} ions of the energy 120 keV and the dose $5 \cdot 10^{13}$ ions/cm^2, C11 – p-type silicon sample implanted with Ar^{+8} ions of the energy 120 keV and the dose $5 \cdot 10^{14}$ ions/cm^2 and C12 – p-type silicon sample implanted with Ar^{+8} ions of the energy 120 keV and the dose $5 \cdot 10^{15}$ ions/cm^2

Discussion

For the pumping beam of light of the wavelength 660 nm the optical absorption coefficient for silicon is β=2630 cm^{-1}. The lifetimes of carriers for n-type sample nonimplanted sample was τ=18 μs. The lifetime of carriers for p-type nonimplanted sample was τ=15 μs.

When we assume that the lifetime of carriers in the damaged layer is much shorter than in the substrate then the decrease of the amplitude of the MFCA signal is expected according to the formula given below:

$$S = \int_{d}^{l} \beta \cdot \exp(-\beta \cdot x) \cdot \partial x \tag{1}$$

When the parameter k is the ratio of the amplitude of the MFCA signal of implanted silicon for the frequency in the range 10^3 Hz – 10^4 Hz to that of the nonimplanted signal when the thickness of the damaged layer can be computed as:

$$d = \frac{\ln(k)}{\beta} \tag{2}$$

For n type silicon samples the following parameters k and d were obtained: k_1=0.059 – d_1=10 μm, k_2=0.04 – d_2=12 μm (see Fig.2).

For p-type silicon samples the following parameters k and d were obtained: k_1=0.017 – d_1=15 μm, k_2=0.005 – d_2=20 μm and k_2=0.003 – d_2=22 μm (see Fig.3).

Such thicknesses of damaged layers were also measured with an argon laser of the wavelength 514 nm where the optical absorption coefficient was 6300 cm^{-1}. Obtained thicknesses for this wavelength of the pumping light was almost identical with these presented in this paper what confirmed the theoretical model taken for computations of the thicknesses of damaged layers being the result of the high energy ion implantation.

The decrease of the amplitude of the MFCA signal in the frequency range from 1 Hz to 10^3 Hz is caused by the diffusion of carriers from the damaged layer to the substrate. With the increase of the frequency of modulation the diffusion length of carriers decreases and fewer carriers diffuse to the region of the substrate. For high frequencies of modulation the MFCA signal comes only from the substrate not from the implantation damaged layer.

Conclusions

The preliminary experimental results of the MFCA measurements of nonimplanted and implanted silicon samples, presented in the paper, indicate that the implantation process considerably modifies the frequency MFCA characteristics. The observed strong decrease of the amplitude of the MFCA signal was interpreted in a model of a damaged layer created in the result of a high energy implantation. Estimated thickness of damaged layers was in the range from 10 μm to 22 μm.

References

1. M. Maliński, Ł. Chrobak, “Determination of the Life Time of Excess Carriers in Silicon With Photoacoustic And Photocurrent Methods”, Journal of Physics: Conference Series 214 (2010), 21-25.
2. Ł. Chrobak, M. Maliński, „Badania parametrów rekombinacyjnych materiałów krzemowych z wykorzystaniem nieniszczącej techniki MFCA opartej na zjawisku modulacji absorpcji na nośnikach swobodnych”, Zeszyty Naukowe Wydziału Elektroniki i Informatyki 3 (2011), WPK, 39-47.
3. Ł. Chrobak, M. Maliński, „Zastosowanie zjawiska modulacji absorpcji na nośnikach swobodnych do nieniszczących badań materiałów półprzewodnikowych”, Elektronika - technologie, konstrukcje, zastosowania LIII (12) (2012), 110-113.

4. Dietzel D., Gibkes J., Chotikaprakhan S., Bein B. K., Pelz J., "Radiometric analysis of laser modulated IR properties of semiconductors", International Journal of Thermophysics 24(3) (2003), 741-755.
5. Li B., Li X., Li W., Huang Q., Zhang X., Accurate determination of electronic transport properties of semiconductor wafers with spatialy resolved photo-carrier techniques, „Journal of Physics: Conference Series", 214 (2010), 012013.
6. Zhang X., Li B., Gao C., Analysis of free carrier absorption measurement of electronic transport properties of silicon wafers, "European Physics Journal Special Topics", 153 (2008), 279-281.
7. Huang Q., Li B., Liu X., Influence of probe beam size on signal analysis of modulated free carrier absorption technique, "Journal of Physics: Conference Series", 214 (2010), 012084.
8. Li W., Li B., Analysis of modulated free-carrier absorption measurement of electronic transport properties of silicon wafers, "Journal of Physics: Conference Series", 214 (2010), 012116.

Abstract

This paper presents experimental amplitude MFCA characteristics of the implanted silicon samples. Influence of the silicon implantation process for frequency MFCA characteristics has been analyzed. The idea and the experimental set-up of the proposed method have been presented and discussed. This paper proves that is possible to estimate depth of the implanted layer from MFCA experimental data.

Streszczenie

W artykule przedstawiono eksperymentalne charakterystyki MFCA dla próbek implantowanego krzemu. Przeanalizowano wpływ procesu implantacji krzemu na charakterystyki MFCA. Przedstawiono i poddano dyskusji ideę zaproponowanej metody oraz stanowisko eksperymentalne. Praca udowadnia, że możliwa jest estymacja głębokości warstwy implantowanej na podstawie eksperymentalnych charakterystyk MFCA.

Anna Witenberg
Maciej Walkowiak
Wydział Elektroniki i Informatyki
Politechnika Koszalińska
ul. Śniadeckich 2, 75-453 Koszalin

Odpowiedzi prądowe w układzie dwóch anten liniowych ze szczególnym uwzględnieniem redukcji późnoczasowych niestabilności

Słowa kluczowe: anteny liniowe, całkowe równanie pola elektrycznego, metoda postępu w czasie (MOT), równanie Hallena w dziedzinie czasu

1. Wprowadzenie

Technologia szerokopasmowa (UWB) jest atrakcyjnym rozwiązaniem w nowoczesnych systemach komunikacji bezprzewodowej. Charakteryzuje ją, między innymi, emisja i odbiór ciągu modulowanych ultrakrótkich impulsów o czasie trwania rzędu kilku nanosekund. Ma to bezpośrednie przełożenie na obniżenie kosztów w układzie nadawanie - odbiór, zmniejsza prawdopodobieństwo zniekształceń transmitowanych informacji oraz ogranicza do minimum możliwość zakłócania procesu przesyłania danych.

Jest wiele dziedzin z możliwością zastosowania tej formy komunikacji:

- technika radiowa, telewizyjna i radarowa,
- szeroko rozumiane obrazowanie w medycynie,
- systemy ochrony i bezpieczeństwa,
- techniki sterowania i identyfikacji [20, 21].

Ze względu na bardzo szerokie pasmo częstotliwości transmitowanych sygnałów, prowadzone obecnie prace koncentrują się na modelowaniu i badaniu właściwości układów antena nadawcza – antena odbiorcza w dziedzinie czasu. Taki stan rzeczy jest naturalną konsekwencją analizy i zachowania się pól elektromagnetycznych, które przecież są procesami, a więc naturalnymi funkcjami czasu. Jest to szczególnie ważne przy projektowaniu i określaniu właściwości systemów antenowych, a szerzej – kanałów komunikacyjnych [22].

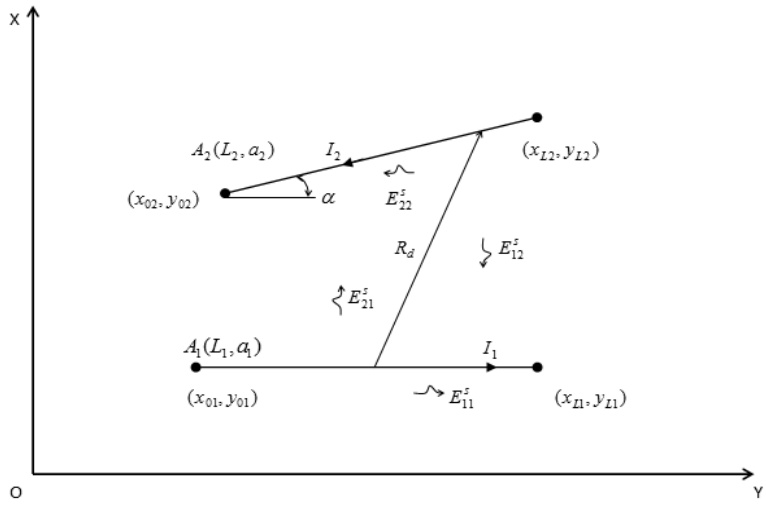
W artykule przedstawiono wyniki modelowania w dziedzinie czasu układu dwóch anten liniowych oświetlonych impulsem pola elektrycznego o kształcie krzywej Gaussa.

Zbadano wpływ długości kroków czasowych przyjętych w obliczeniach oraz wzajemnego położenia anten na wielkość późnoczasowych niestabilności powstających w procesie modelowania [10, 12].

Punktem wyjścia analizy tych zjawisk jest układ całkowych równań Hallena, których rozwiązanie opiera się na odpowiedniej dyskretyzacji w czasie i w przestrzeni. W procesie rozwiązywania równań wykorzystano powszechnie stosowaną w numerycznej analizie zjawisk elektromagnetycznych metodę MOT (marching-on in time) oraz schemat Bubnova-Galerkina [25].

2. Model badanego układu dwóch anten liniowych

Układ poddany modelowaniu w dziedzinie czasu składa się z dwóch doskonale przewodzących anten liniowych A_1 i A_2 o długościach i promieniach odpowiednio równych: (L_1, a_1) i (L_2, a_2), przy czym $a_1 << L_1$ i $a_2 << L_2$, co spełnia założenie przybliżenia cienkoprzewodowego. Położenie anten w przestrzeni może być dowolne, ale dla większej przejrzystości przyjęto dwuwymiarowy układ współrzędnych XY z anteną A_1 położoną równolegle do osi OX. Modelowany układ pokazano na rysunku 1.



Rys. 1. Model badanego układu dwóch anten liniowych; $A_1(L_1, a_1)$ - antena A_1 o długości L_1 i promieniu a_1; $A_2(L_2, a_2)$ - antena A_2 o długości L_2 i promieniu a_2; odległość R_d opisana zależnością (3)

Antena A_1 oświetlona jest impulsem pola elektrycznego o natężeniu $\mathbf{E}_1^i$, pochodzącego z lokalnego generatora umieszczonego w środku anteny. Pola elektryczne w układzie spełniają następujące równania

$$\begin{aligned} &\mathbf{1}_{n1} \times (\mathbf{E}_1^i + \mathbf{E}_{11}^s + \mathbf{E}_{12}^s) = 0 \\ &\mathbf{1}_{n2} \times (\mathbf{E}_{22}^s + \mathbf{E}_{21}^s) = 0 \end{aligned} \tag{1}$$

w których:

$\mathbf{1}_{n1} \cdot$ i $\mathbf{1}_{n2} \cdot$ - wektory normalne do powierzchni anteny pierwszej i drugiej

$\mathbf{E}_1^i$ - pole elektryczne lokalnego generatora pobudzającego antenę A_1 w środku,

$\mathbf{E}_{11}^s$ - składowa pola elektrycznego na antenie A_1 pochodzącego od prądu na A_1,

$\mathbf{E}_{12}^s$ - składowa pola elektrycznego na antenie A_1 pochodzącego od prądu na A_2,

$\mathbf{E}_{22}^s$ - składowa pola elektrycznego na antenie A_2 pochodzącego od prądu na A_2,

$\mathbf{E}_{21}^s$ - składowa pola elektrycznego na antenie A_2 pochodzącego od prądu na A_1.

Impulsową (w dziedzinie czasu) odpowiedź modelowanych anten opisuje następujący układ równań Hallena [19, 25]:

$$\begin{aligned} &\int_{x_{01}}^{x_{L1}} \frac{I_1(x', t - R_{a1}/c)}{4\pi R_{a1}} dx' + \int_{x_{02}}^{x_{L2}} \frac{I_2(x', t - R_d/c - \tau)\cos\alpha}{4\pi R_d} dx' = \\ &= F_{01}(t - \frac{x - x_{01}}{c}) + F_{L1}(t - \frac{x_{L1} - x}{c}) + \\ &+ \frac{1}{2\eta_0} \int_{x_{01}}^{x_{L1}} E_1^i(x', t - \frac{|x - x'|}{c}) dx' \\ &\int_{x_{01}}^{x_{L1}} \frac{I_1(x', t - R_d/c - \tau)\cos\alpha}{4\pi R_d} dx' + \int_{x_{02}}^{x_{L2}} \frac{I_2(x', t - R_{a2}/c)}{4\pi R_{a2}} dx' = \\ &= F_{02}(t - \frac{x - x_{02}}{c}) + F_{L2}(t - \frac{x_{L2} - x}{c}) \end{aligned} \tag{2}$$

w którym:

I_1 i I_2 są natężeniami prądów płynących odpowiednio w antenach 1 i 2,

η_0 - impedancja falowa próżni; $\eta_0 = 120\pi$,

c - prędkość fali elektromagnetycznej w próżni; $c = 3 \cdot 10^8\ m/s$.

Odległości między punktami obserwacji a punktami całkowania na antenach dane są wzorami:

$$R_{a1} = \sqrt{(x - x_1')^2 + a_1}$$
$$R_{a2} = \sqrt{(x - x_2')^2 + (y - y_2')^2 + a_2^2} \tag{3}$$
$$R_d = \sqrt{(x - x')^2 + (y - y')^2}.$$

Czas opóźnienia τ w oświetlaniu elementów x' jest równy

$$\tau = \frac{x' \cdot \sin\alpha}{c} \tag{4}$$

Funkcje $F_{Om}(t)$ i $F_{Lm}(t)$ $(m = 1,2)$ opisujące wielokrotne odbicia fal prądowych na końcach anten wyrażają się przez pewne funkcje pomocnicze $K_{Om}(t)$ i $K_{Lm}(t)$ zdefiniowane w [23, 25], następująco:

$$F_{Om}(t) = \sum_{n=0}^{\infty} K_{Om}\left(t - \frac{2 \cdot n \cdot L_m}{c}\right) - \sum_{n=0}^{\infty} K_{Lm}\left(t - \frac{(2n+1) \cdot L_m}{c}\right) \tag{5}$$

$$F_{Lm}(t) = \sum_{n=0}^{\infty} K_{Lm}\left(t - \frac{2 \cdot n \cdot L_m}{c}\right) - \sum_{n=0}^{\infty} K_{Om}\left(t - \frac{(2n+1) \cdot L_m}{c}\right) \tag{6}$$

i wyznaczane są dla warunków granicznych na końcach anten

$$I_m((x_{Om}, y_{Om}); t) = I_m((x_{Lm}, y_{Lm}); t) = 0.$$

Zakładamy również, że anteny nie są wzbudzone przed czasem $t = 0$, tzn. $I_m(t) = 0$ dla $t \le 0$.

Układ równań (2) rozwiązujemy metodą momentów [12] zamieniając go na układ równań macierzowych.

Prąd $I_m(x,t)$ w równaniach (2) aproksymujemy zależnością [24]

$$I_m(x,t) = \sum_{n=1}^{N} I_{m,n}(t_k) \cdot f_{m,n}(x) \qquad (m = 1,2) \tag{7}$$

w której $I_{m,n}(t_k)$ jest nieznaną wartością natężenia prądu na *n*-tym segmencie *m*-tej anteny, a $f_{m,n}(x)$ jest pewną przestrzenną funkcją bazową. Innymi słowy, rozwiązując numerycznie układ równań (7) do dyskretyzacji przestrzeni wykorzystujemy procedurę Galerkina-Bubnova oraz schemat MOT do dyskretyzacji w czasie.

3. Specyfikacja założeń przyjętych do obliczeń numerycznych

Analizie poddano stabilność późnoczasowej odpowiedzi prądowej anten, tzn. wartości prądów zaindukowanych na antenach w funkcji długości (wartości) kroków czasowych Δt dla anteny A_1 i wartości kątów α pod jakimi impuls pola elektrycznego pobudza antenę A_2.

Obliczenia przeprowadzono dla trzech wartości Δt:

- $\Delta t_1 = \dfrac{\Delta x}{c} = 0{,}1667 ns$,
- $\Delta t_2 = 2 \cdot \dfrac{\Delta x}{c} = 0{,}3334 ns$,
- $\Delta t_3 = 2{,}5 \cdot \dfrac{\Delta x}{c} = 0{,}4168 ns$

oraz dwóch wartości α:

- $\alpha_1 = 30^\circ$,
- $\alpha_2 = 60^\circ$.

Odpowiedzi prądowe na antenach pokazano w punktach będących ich geometrycznym środkiem i odległych od siebie o $d = 2m$.

Przyjęto następujące parametry geometryczne anten:

- $2L_1 = 2L_2 = 1m$,
- $a_1 = a_2 = 0{,}002m$.

Każdą z anten podzielono na 20 segmentów o jednakowej długości $\Delta x = 0{,}05m$. Antenę A_1 pobudzono w środku impulsem pola elektrycznego E_1^i o kształcie krzywej Gaussa

$$E_1^i = \frac{4E_0}{cT\sqrt{\pi}} \exp\left[-\left(\frac{4}{T}\left(t - t_0\right) \right)^2 \right] \tag{8}$$

gdzie:

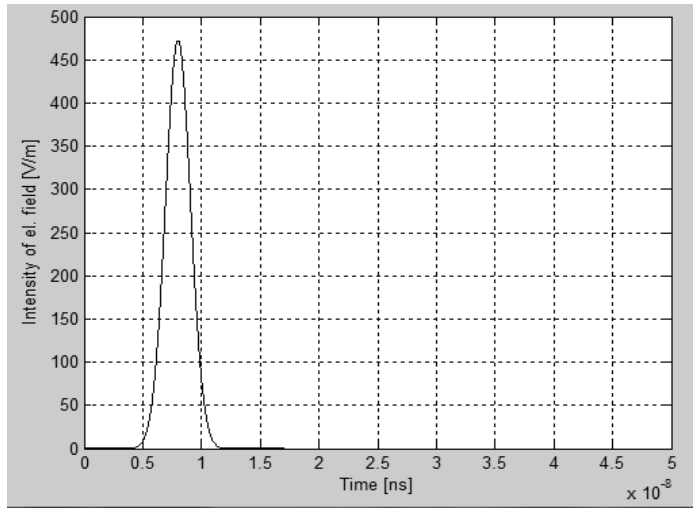
E_0 - wartość natężenia pola elektrycznego; $E_0 = 120\pi$,

T - szerokość impulsu; $T = 6\ ns$,

t_0 - przesunięcie maksimum impulsu; $t_0 = 8\ ns$,

c - prędkość fali elektromagnetycznej; $c = 3 \cdot 10^8\ m/s$.

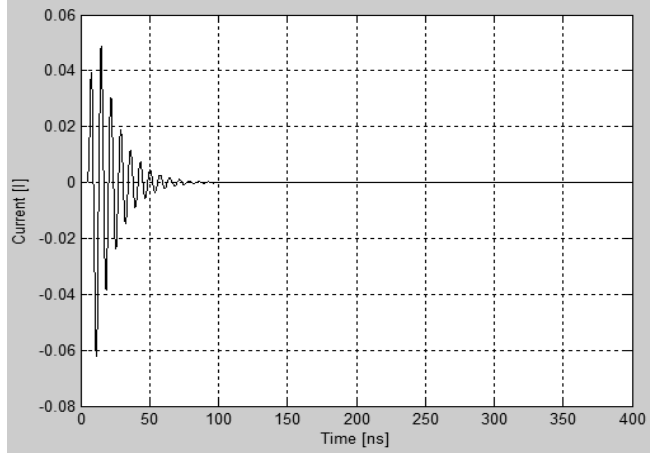
Przebieg impulsu Gaussa (8) pokazano na rysunku 2.



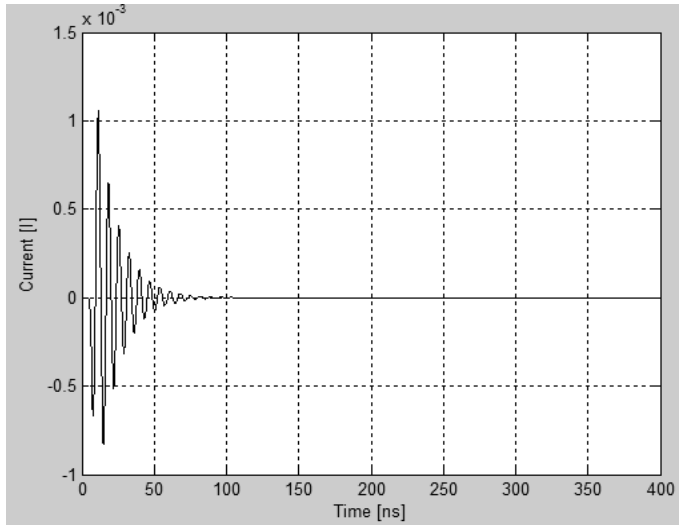
Rys. 2. Impuls fali elektromagnetycznej E_1^i opisany równaniem (8)

4. Symulacje komputerowe i analiza wyników

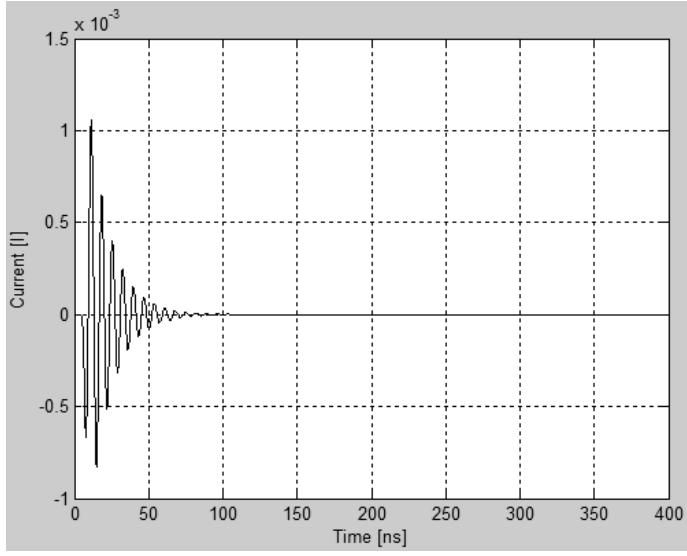
Na rysunkach 3-11 przedstawiono wyniki symulacji komputerowych, tzn. prądy w środku anten A_1 i A_2 w modelu jak na rysunku 1. Dla kroku czasowego $\Delta t_1 = 0{,}1667\ ns$:



Rys. 3. Odpowiedź prądowa anteny A_1

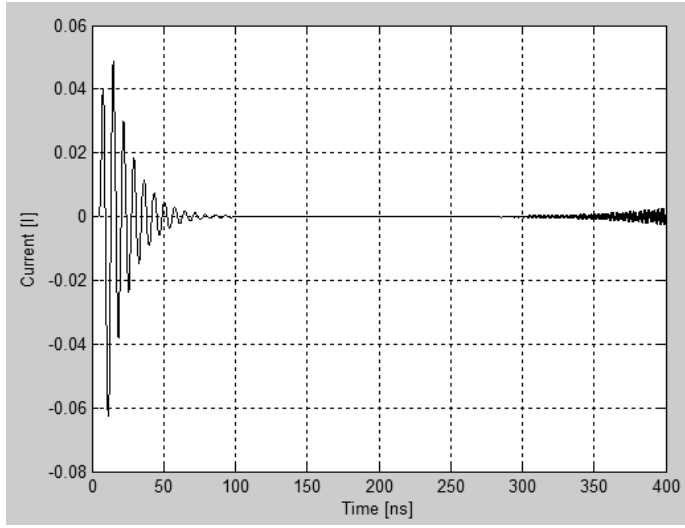


Rys. 4. Odpowiedź prądowa anteny A_2 dla kąta $\alpha = 30^\circ$

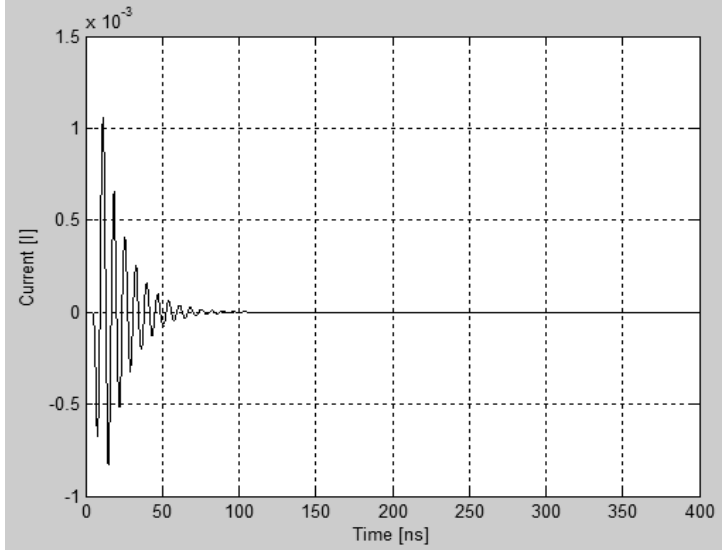


Rys. 5. Odpowiedź prądowa anteny A_2 dla kąta $\alpha = 60^\circ$

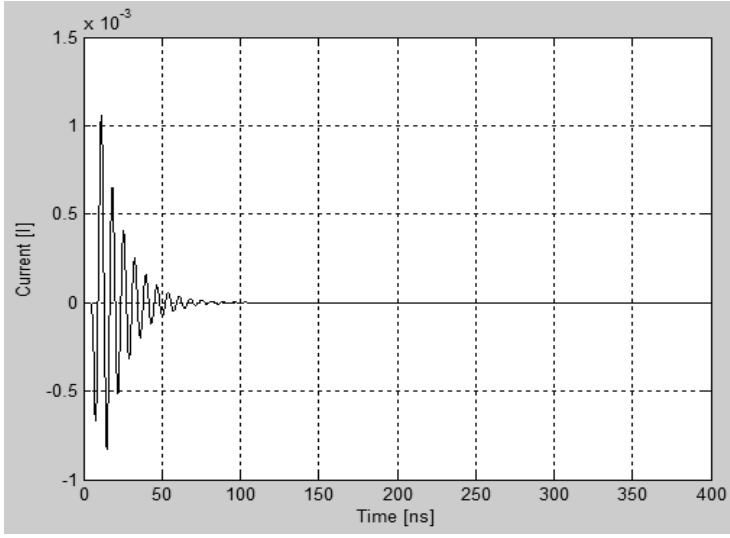
Dla kroku czasowego $\Delta t_2 = 0{,}3334\ ns$:



Rys. 6. Odpowiedź prądowa anteny A_1

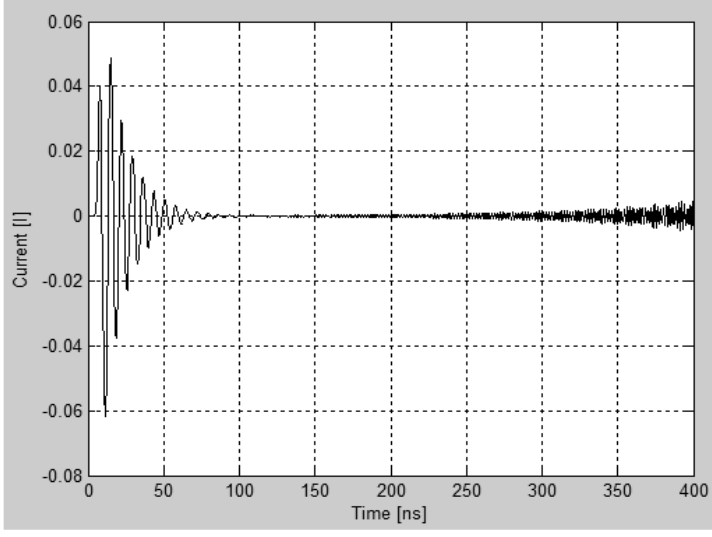


Rys. 7. Odpowiedź prądowa anteny A_2 dla kąta $\alpha = 30^\circ$

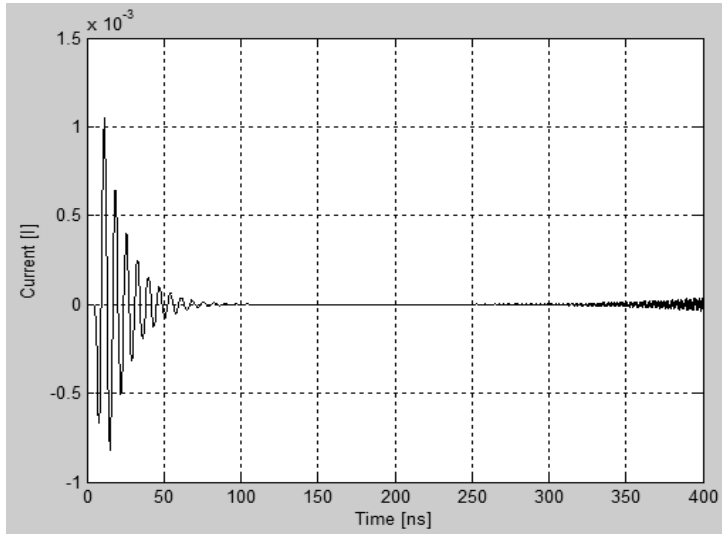


Rys. 8. Odpowiedź prądowa anteny A_2 dla kąta $\alpha = 60^\circ$

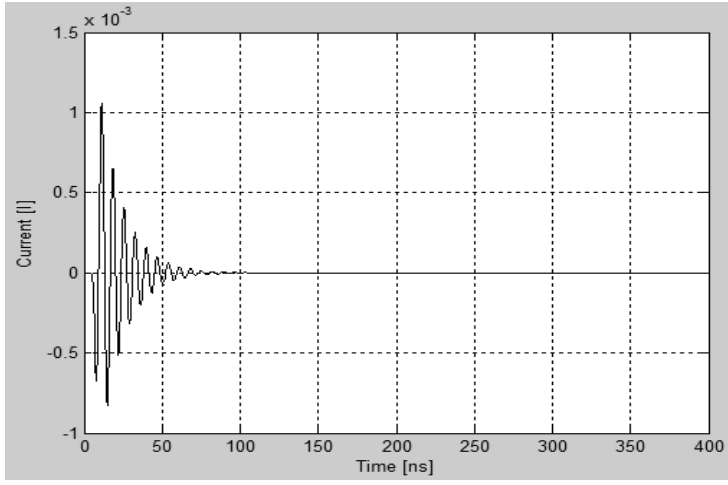
Dla kroku czasowego $\Delta t_3 = 0{,}4168\ ns$:



Rys. 9. Odpowiedź prądowa anteny A_1



Rys. 10. Odpowiedź prądowa anteny A_2 dla kąta $\alpha = 30^\circ$



Rys. 11. Odpowiedź prądowa anteny A_2 dla kąta $\alpha = 60^\circ$

W tabeli 1 zebrano wyniki symulacji komputerowych. Uwzględniono w niej późnoczasowe odpowiedzi anten pod względem pojawienia się niestabilności.

Tabela 1. Wyniki symulacji komputerowych dla przyjętego modelu anten liniowych, „-" oznacza brak niestabilności w późnoczasowej odpowiedzi anteny, „+" oznacza pojawienie się niestabilności w późnoczasowej odpowiedzi anteny

Antena					
A_1		A_2			
		$\alpha = 30^\circ$		$\alpha = 60^\circ$	
krok czasowy	niestabilności	krok czasowy	niestabilności	krok czasowy	niestabilności
0,1667 *ns*	-	0,1444 *ns*	-	0,0834 *ns*	-
0,3334 *ns*	+	0,2887 *ns*	-	0,1667 *ns*	-
0,4168 *ns*	+	0,3609 *ns*	+	0,2084 *ns*	-

5. Podsumowanie

Wyniki badań odpowiedzi prądowych w układzie dwóch anten liniowych prowadzą do wniosków:

- brak niestabilności w późnoczasowej odpowiedzi anteny bezpośrednio oświetlonej impulsem Gaussa (antena A_1) dla długości kroku czasowego $\Delta t = \Delta x / c$ i obecność niestabilności dla $\Delta t > \Delta x / c$,
  - dla $\Delta t = 2\Delta x / c$ po około $300\ ns$,
  - dla $\Delta t = 2{,}5\Delta x / c$ po około $150\ ns$,
- brak niestabilności w późnoczasowej odpowiedzi anteny A_2 dla długości kroków czasowych $\Delta t = \Delta x / c$ i $\Delta t = 2\Delta x / c$ (antena A_1) i wartościach kąta $\alpha = 30^\circ$ i $\alpha = 60^\circ$,
- obecność niestabilności w odpowiedzi anteny A_2 dla długości kroku czasowego $\Delta t = 2{,}5\Delta x / c$ (antena A_1) i $\alpha = 30^\circ$ (po około $300\ ns$),
- brak niestabilności w odpowiedzi anteny A_2 dla długości kroku czasowego $\Delta t = 2{,}5\Delta x / c$ (antena A_1) i $\alpha = 60^\circ$.

Fakt pojawienia się oscylacji (lub ich brak) w późnoczasowej odpowiedzi anteny A_2 jest silnie skorelowany z długością kroku czasowego przyjętą w wyznaczaniu odpowiedzi prądowej anteny A_1 oraz wartością kąta α. Kąt α wpływa na zmniejszenie długości kroku czasowego przy wyznaczaniu odpowiedzi prądowej anteny A_2 (zależność odwrotnie proporcjonalna). Na skuteczniejsze tłumienie oscylacji w późnoczasowej części rozwiązania układu równań (2) metodą MOT może istotnie wpływać wartość kąta α zmniejszająca błędy systematyczne wprowadzane do rozwiązania. Udowodnienie tej tezy wymaga dalszych badań.

W uzupełnieniu dodajmy, że każda z odmian metody MOT ma swoje wady i zalety i nie jest wolna od niestabilności mających charakter narastających oscylacji o wysokiej częstotliwości, które pojawiają się w późnoczasowej części rozwiązania. Dokładność i stabilność rozwiązań całkowych równań Hallena w dziedzinie czasu silnie zależy od długości (wartości) kroku czasowego Δt przyjętego w procesie obliczeń.

Bibliografia

1. Rao S. M.: *Time Domain Electromagnetics*, Academic Press, London, 1999.
2. Jung B. H., Sarkar T. K.: *Time-domain electric-field integral equation with central finite difference*, Microwave and Optical Technology Letters, 2001, vol.31, no.6, pp. 429-434.
3. Manara G.: *A space-time discretization criterion for a stable time-marching solution of the electric field integral equation*, IEEE Transactions on Antennas and Propagation, 1997, vol.45, no.3, pp. 527-532.
4. Weile S. D., Pisharody G., Chen N. Y., Shanker B., Michielssen E.: *A novel scheme for the solution on the time-domain integral equations of electromagnetics*, IEEE Transactions on Antennas and Propagation, 2004, vol.52, no.1, pp. 283-295.
5. Chung Y. S.: *Solution of time domain electric field integral equation using the Laguerre polynomials*, IEEE Transactions on Antennas and Propagation, 2004, vol.52, no.9, pp. 2319-2328.
6. Vechinski A. D., Rao S. M.: *A stable procedure to calculate the transient scattering by conducting surfaces of arbitrary shapes*, IEEE Transactions on Antennas and Propagation, 1992, vol.40, no.6, pp. 661-665.
7. Sadigh A., Arvas E.: *Treating the instabilities in marching-on-in-time method from a different perspective*, IEEE Transactions on Antennas and Propagation, 1993, vol.41, no.12, pp. 1695-1702.
8. Wang X., Wildman R. A., Weile D. S., Monk P. A.: *A finite difference delay modeling approach ito the discretization of the time domain integral equations of electromagnetics*, IEEE Transactions on Antennas and Propagation, 2008, vol.56, no.8, pp. 2442-2452.

9. Davis P. J.: *On the stability of time-marching schemes for the general surface electric field integral equation*, IEEE Transactions on Antennas and Propagation, 1996, vol.44, no.11, pp. 1467-1473.
10. Makarov S. N.: *Antenna and EM Modeling with Matlab,* John Wiley & Sons, New York, 2002.
11. Rao S. M., Wilton D. R., Glisson A. W.: *Electromagnetic scattering by surfaces of arbitrary shape*, IEEE Transactions on Antennas and Propagation, 1982, vol.30, no.5, pp. 409-418.
12. Jung B. H., Sarkar T. K., Ji Z., Chung Y.: *Time- domain analysis of conducting wire antennas and scatterers*, Microwave and Optical Technology Letters, 2003, vol.38, no.6, pp. 433-436.
13. Rao S. M., Sarkar T. K.: *An efficient method to evaluate the time-domain scattering from arbitrarily shaped conducting bodies*, Microwave and Optical Technology Letters, 1998, vol.17, pp. 321-325.
14. Liu T. K., Mei K. K.: *A time domain integral equation solution for linear antennas and scatterers*, Radio Sci., 1973, vol.8.
15. Smith P. D.: *Instabilities in time marching methods for scattering: cause and rectification*, Electromagn., 1990, vol.10, pp. 439-451.
16. Walkowiak M.: *Zjawiska przejściowe w antenach liniowych i rozpraszaczach*, WSP, Zielona Góra, 1994.
17. Witenberg A., Walkowiak M.: *Rozkład prądu wzdłuż anten liniowych – modele Pocklingtona i Hallena*, Zeszyty Naukowe Wydziału Elektroniki i Informatyki Politechniki Koszalińskiej, 2010, nr 2, s. 37-46.
18. Witenberg A., Walkowiak M.: *Wykorzystanie wielomianów Laguerre'a do rozwiązania równania Hallena w dziedzinie czasu*, Metody Informatyki Stosowanej, Szczecin, 2011, nr 1, s. 183-192.
19. Sagnard F., Uguen B., El-Zein G.: *Reception of an oblique electromagnetic plane wave by a linear-wire antenna: A time domain analysis*, Microwave and Optical Technology Letters, 2003, vol.38, no.4, pp. 281-291.
20. Sagnard F., El-Zein G.: *Waveform prediction of a pulse Communications link between antennas modeled by a combination of thin-wires*, Progress in Electromagnetics Research Symposium, 2006, Cambridge, USA, pp. 137-142.
21. Smith G.: *Teaching antenna reception and scattering from a time – domain prespective*, AM. J. Phys., 2002, vol.70, pp. 829-844.
22. Immoreev I. J.: *Radiation of ultra-wideband (UWB) signals*, Radio Physics and Radio Astronomy, 2002, vol.7, no.4, pp. 389-393.
23. Gomez Martin R., Rubio Bretones A., Gonzales Garcia S.: *Some thoughts about transient radiation by straight thin wires*, IEEE Transactions on Antennas and Propagation Magazine, 1999, vol.41, no.3, pp. 24-33.

24. Witenberg A., Walkowiak M.: *Rozwiązanie równania pola elektrycznego modelującego anteny liniowe w dziedzinie czasu*, Zeszyty Naukowe Wydziału Elektroniki i Informatyki Politechniki Koszalińskiej, 2011, nr 3, s. 123-132.
25. Poljak D., Antonijevic S., Drissi K., Kerroum K.: *Transient Response of Straight Thin Wires Located at Different Heights Above a Ground Plane Using Antenna Theory and Transmission Line Approach*, IEEE Transactions on Electromagnetic Compatibility, 2010, vol.52, no.1, pp. 108-115.

Streszczenie

W artykule przedstawiono wyniki modelowania w dziedzinie czasu układu dwóch anten liniowych za pomocą układu całkowych równań Hallena. Do wyznaczenia odpowiedzi prądowych anten wykorzystano powszechnie stosowaną metodę postępu w czasie (metoda MOT) oraz schemat Galerkina-Bubnova. Wyniki modelowania potwierdzają słuszność przyjętych założeń o bezpośrednim wpływie długości kroków czasowych przyjętych w obliczeniach i położenia anten względem siebie na wielkość niestabilności pojawiających się w późnoczasowych odpowiedziach prądowych badanych obiektów.

Abstract

This article is focused on modeling of two interacting thin-wire antennas in time domain. The time-domain antenna theory formulation is based on a set of the space-time Hallen integral equations. We used the marching-on in time (MOT) method and Galerkin-Bubnov procedure to obtain a transient responses of antennas. An analytic formulation to study the reception process, at an oblique angle of incidence, of a wire antennas excited by a transient electric fields (short Gaussian pulses) is proposed. The length of time steps and angles of incidence influences the stability of the MOT algorithm. These facts are discussed and demonstrated in numerical results.

Keywords: linear antennas, electric field integral equation (EFIE), solution of time domain EFIE – late time instabilities, marching-on-in time method (MOT), Hallen's equation in time domain

Marcin Walczak
Katedra Systemów Elektronicznych
Wydział Elektroniki i Informatyki
Politechnika Koszalińska

Badania symulacyjne charakterystyk przetwornic buck i boost z uwzględnieniem rezystancji pasożytniczych

Słowa kluczowe: BUCK, BOOST, przetwornica, Step-down, Step-up, charakterystyki przetwornic, rezystancje pasożytnicze

1. Wstęp

Ciągły rozwój techniki, tworzenie coraz bardziej złożonych i uniwersalnych układów scalonych, a także gwałtowny rozwój urządzeń mobilnych doprowadziły do tego, że coraz większą wagę zaczęto przykładać do ilości zużywanej energii elektrycznej. Pojęcie energooszczędności dotyczy energii pobieranej przez urządzenie, czyli sumy energii wykorzystywanej oraz energii traconej np. podczas zmniejszania, podwyższania lub stabilizacji napięcia. Dodatkowo coraz nowocześniejsze i szybsze układy scalone potrzebowały do prawidłowej pracy coraz mniejszych napięć i coraz większych prądów. To wszystko spowodowało, że wszędzie, gdzie to było możliwe zaczęto stosować układy przetwornic, które charakteryzują się wysoką sprawnością oraz niewielkimi wymiarami w stosunku do ilości przetwarzanej energii. Z drugiej strony przetwornica jest układem nieliniowym, działającym w sposób impulsowy więc jej modelowanie jest bardziej skomplikowane niż w przypadku układów liniowych, pracujących w sposób ciągły. Patrząc na układ przetwornicy pod kątem zasady działania można w niej wyróżnić dwa stany – stan włączenia (ON) oraz wyłączenia (OFF). Stwierdzenie, że przetwornica znajduje się w stanie włączenia oznacza, że jej klucz tranzystorowy jest zwarty i przewodzi prąd elektryczny. Suma czasów ON i OFF nazywana jest okresem kluczowania. Stosunek czasu włączenia (ON) do całego okresu kluczowania nazywa się współczynnikiem wypełnienia d_A. Ponieważ przetwornice pracują impulsowo, to ich modele matematyczne odnoszą się do wartości uśrednionych w stosunku do całego okresu kluczowania. Jedną z możliwości wyznaczenia takiego modelu jest napisanie równań dla obu stanów ON i OFF, korzystając z ogólnej teorii obwodów. Następnie równania te należy uśrednić w odniesieniu do całego okresu kluczowania [1][2]. Przy pomocy linearyzacji oddziela się wielkości wielkosygnałowe (niezmienne w czasie) od małosygnałowych (zmiennych w czasie). Uzyskane w ten sposób wzory stanowią

komplet równań opisujących działanie przetwornicy dla dostatecznie małych częstotliwości (tj. poniżej 0,1 częstotliwości kluczowania) [3][4].

Na podstawie matematycznego modelu przetwornicy wyznacza się jej charakterystyki. Najczęściej sprawdzane jest zachowanie amplitudy danego napięcia lub prądu w dziedzinie czasu oraz częstotliwości. Charakterystyka w dziedzinie czasu ma za zadanie odpowiedzieć na pytanie jak zachowa się dane napięcie/prąd po skokowej zmianie np. napięcia zasilania. Zbyt duże wahania sygnału badanego są w niektórych przypadkach wysoko niepożądane. Charakterystyka w dziedzinie częstotliwości pozwala uzyskać informacje o przetwornicy jako filtrze dolnoprzepustowym. W tym przypadku ważna jest szerokość pasma przepustowego, wzmocnienie, tłumienie w paśmie zaporowym, a także przesunięcie fazowe.

W trakcie pracy przetwornicy, w jej elementach zarówno przełączających (tranzystor, dioda) jak i pasywnych (cewka kondensator) generowane są straty mocy. Straty te są wynikiem obecności rezystancji pasożytniczych, które znacząco wpływają na wszystkie charakterystyki przetwornic.

Konieczność uwzględniania tych rezystancji w modelach przetwornic impulsowych zauważono już przy opisie pierwszych technik modelowania [5]. Jednak uwzględniane rezystancje dotyczyły jedynie rezystancji elementów pasywnych (cewki i kondensatora), co na przykład w przypadku przetwornic Ćuk'a prowadziło do błędów przy symulacjach [6]. Model poprawnie opisujący działanie przetwornicy jest podstawą przy projektowaniu bloku sterowania. Blok ten pozwala przetwornicy właściwie reagować na zakłócenia o niskich częstotliwościach (do ok. 0,1 częstotliwości kluczowania). Pomimo tego rezystancje pasożytnicze kluczy (tranzystora i diody) są wciąż pomijane w niektórych pracach [1][7][8]. W pracach [9][10] pokazano, że wpływ tych rezystancji na charakterystyki przetwornicy jest zauważalny i nie powinien być pomijany. Natomiast wpływ poszczególnych rezystancji pasożytniczych na wahania napięcia wyjściowego i prądu wejściowego, w stanie nieustalonym oraz offsetu w stanie ustalonym opisano w pracy [11].

Niniejsza praca jest kontynuacją badań przedstawiających wpływ rezystancji pasożytniczych na charakterystyki częstotliwościowe oraz czasowe przetwornic BUCK oraz BOOST. Przedstawiono w niej symulacje bloku głównego przetwornic, zawierającego rzeczywiste, zmierzone rezystancje pasożytnicze zarówno elementów pasywnych (cewka, kondensator) jak i elementów łączeniowych (tranzystor, dioda). Symulacje zawierają porównanie transmitancji przetwornicy:

- idealnej,
- uwzględniającej rezystancje pasożytnicze elementów biernych,
- uwzględniającej wszystkie rezystancje pasożytnicze.

Drugi rozdział poświęcony jest przetwornicy BUCK. Na początku tego rozdziału przedstawiono definicję oraz wzory na transmitancje $H_g(s)$ oraz $H_d(s)$ wspomnianej przetwornicy. W kolejnych dwóch częściach tego rozdziału przedstawiono

charakterystyki częstotliwościowe oraz symulacje odpowiedzi na uskok jednostkowy, w dziedzinie czasu. Opisano znaczenie tych symulacji.

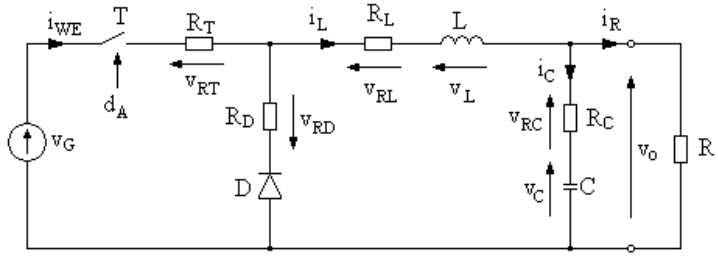
Trzeci rozdział poświęcono przetwornicy BOOST. Podobnie jak w poprzednim przypadku pierwszy podpunkt tego rozdziału zawiera wyrażenia na transmitancje $H_g(s)$ oraz $H_d(s)$, a w dwóch pozostałych zamieszczono charakterystyki częstotliwościowe oraz wyniki symulacji odpowiedzi na uskok jednostkowy, w dziedzinie czasu.

Rozdział czwarty zawiera podsumowanie pracy.

2. Przetwornica BUCK

2.1. Postacie transmitancji Hd oraz Hg

Blok główny przetwornicy BUCK, zawierającej wszystkie rezystancje pasożytnicze, przedstawiono na rys. 1. Do elementów kluczujących zalicza się tranzystor przedstawiony jako łącznik „T" oraz diodę „D". Odpowiadające im rezystancje pasożytnicze, to odpowiednio R_T oraz R_D. Pozostałe rezystancje pasożytnicze R_L oraz R_C dotyczą strat w cewce i kondensatorze. Sygnałem sterującym pracą tranzystora jest sygnał d_A.



Rys. 1. Blok główny przetwornicy BUCK zawierający wszystkie rezystancje pasożytnicze

W tabeli 1 zamieszczono parametry symulacji. Pojemność kondensatora, indukcyjność cewki oraz wszystkie rezystancje pasożytnicze, wraz z rezystancją obciążenia, zostały określone w drodze pomiaru i odnoszą się do elementów przetwornicy, wykonanej w rzeczywistości.

Tabela 1. Parametry elementów bloku głównego przetwornicy BUCK

V_G=12 V	R=5 Ω
D_A=0,5	R_T=28 mΩ
L=92,2 uH	R_D=300 mΩ
C=487,23 uF	R_L=40,1 mΩ
	R_C=42,8 mΩ

Wpływ zmian napięcia wejściowego na napięcie wyjściowe opisuje transmitancja $H_g(s)$. Transmitancję tę określa się jako stosunek małosygnałowej wartości napięcia wyjściowego $V_o(s)$ do małosygnałowej wartości napięcia wejściowego $V_g(s)$ przy stałym współczynniku wypełnienia, zgodnie z poniższym wzorem:

$$H_g(s) = \left.\frac{V_o(s)}{V_g(s)}\right|_{\theta(s)=0} \tag{1}$$

gdzie:

$\theta(s)$ – składowa zmienna współczynnika wypełnienia, w dziedzinie „s"

W przypadku układu z rys. 1 transmitancja $H_g(s)$ wynosi [4]:

$$H_g(s) = \frac{D(1+sCR_C)}{LC_Z s^2 + (GL + R_Z C_Z + CR_C)s + GR_Z + 1} \tag{2}$$

gdzie:

D – składowa stała współczynnika wypełnienia

$$R_Z = D_A(R_T - R_D) + R_D + R_L$$

$$C_Z = C(1 - R_C G)$$

$$G = \frac{1}{R}$$

Gdy układ sterowania przetwornicy wykryje zmianę napięcia wyjściowego spowodowaną np. zmianą napięcia wejściowego, to odpowiednio zareaguje modyfikując współczynnik wypełnienia sygnału kluczującego. Reakcję tę opisuje transmitancja $H_d(s)$, którą rozumie się jako stosunek małosygnałowej wartości napięcia wyjściowego $V_o(s)$ i małosygnałowej wartości współczynnika wypełnienia $\theta(s)$ przy stałej wartości napięcia wejściowego $V_G(s)$, zgodnie ze wzorem:

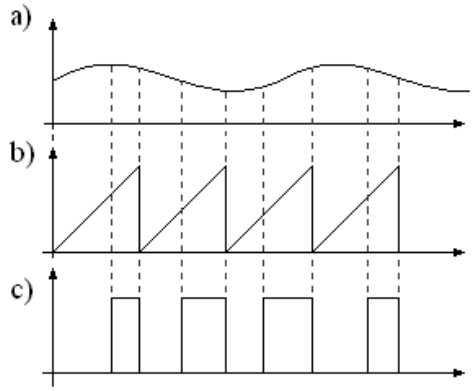
$$H_d(s) = \left.\frac{V_o(s)}{\theta(s)}\right|_{V_g(s)=0} \tag{3}$$

W rzeczywistym układzie skokowa zmiana np. napięcia wejściowego może być przyczyną dużych oscylacji napięcia na wyjściu przetwornicy. Amplituda oraz czas trwania tych oscylacji zależy głównie od typu przetwornicy oraz elementów jej bloku głównego. Szybsze stłumienie oscylacji jest możliwe poprzez zastosowanie pętli sprzężenia zwrotnego z układem sterującym. Jak wspomniano układ ten steruje współczynnikiem wypełnienia tj. czasem włączenia łącznika tranzystorowego.

Sam układ sterujący jest projektowany między innymi na podstawie transmitancji $H_d(s)$, dzięki temu potrafi on prawidłowo zareagować na wszelkie zmiany pojawiające się w napięciu wyjściowym przetwornicy. Źle wyznaczona transmitancja $H_d(s)$ spowoduje, że układ sterowania nie będzie w stanie prawidłowo sterować pracą przetwornicy, co

w ostateczności może doprowadzić do pojawienia się długotrwałych oscylacji napięcia wyjściowego.

W najprostszy sposób reakcję układu sterowania w przetwornicy BUCK można przedstawić za pomocą rys. 2. Napięcie wyjściowe (rys. 2a) podawane jest na komparator wraz z sygnałem piłokształtnym (rys. 2b). Na wyjściu komparatora otrzymuje się sygnał prostokątny o różnym współczynniku wypełnienia. Jeżeli napięcie wyjściowe było zbyt duże, to układ sterowania zmniejsza współczynnik wypełnienia, aby je obniżyć. Jeżeli napięcie na wyjściu było zbyt małe, to współczynnik wypełnienia zostaje zwiększony (rys. 2c).



Rys. 2. Wpływ sygnału wejściowego (a), na kształtowanie współczynnika wypełnienia (c) w przetwornicy BUCK

W przypadku układu z rys. 1 transmitancja $H_d(s)$ wynosi [4]:

$$H_d(s) = \frac{(V_G - I_L(R_T - R_D)) \cdot (1 + sCR_C)}{LC_Z s^2 + (GL + R_Z C_Z + CR_C)s + GR_Z + 1} \tag{4}$$

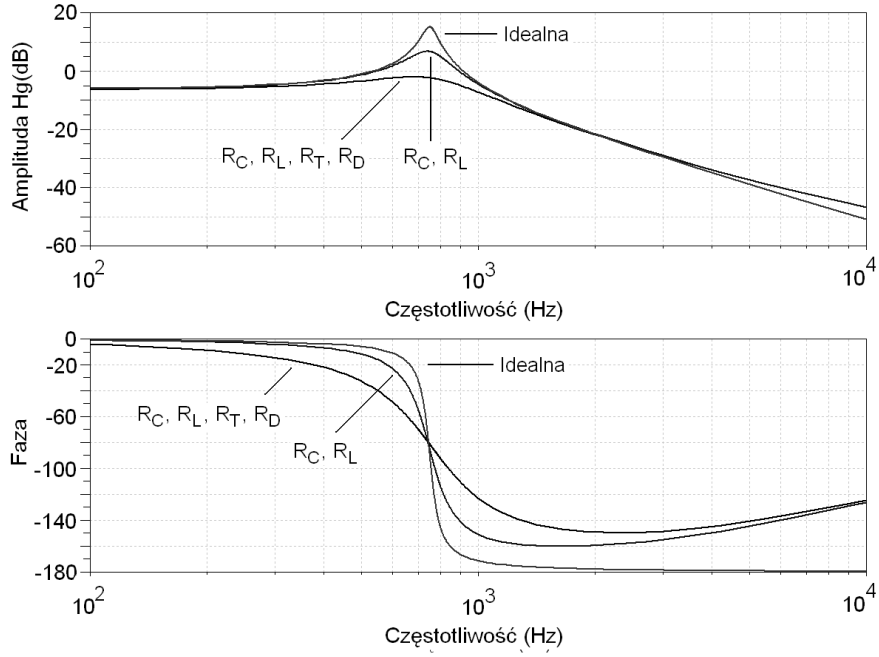
gdzie składowa stała prądu cewki wynosi:

$$I_L = G \frac{D_A V_G}{1 + R_Z G}$$

2.2. Symulacje w dziedzinie częstotliwości

Symulacja transmitancji $H_g(s)$ w dziedzinie częstotliwości pozwala określić jak częstotliwość sygnału wejściowego wpływa na przenoszenie tego sygnału z wejścia na wyjście przetwornicy. W praktyce oznacza to informację o przenoszeniu zakłóceń o różnych częstotliwościach z wejścia na wyjście układu. Na rys. 3 przedstawiono wyniki symulacji amplitudy i fazy transmitancji $H_g(s)$. Wyniki dotyczą trzech przypadków: przetwornicy idealnej, przetwornicy z rezystancjami pasożytniczymi elementów pasywnych (R_C, R_L) oraz przetwornicy z uwzględnieniem wszystkich

rezystancji pasożytniczych (R_C, R_L, R_T, R_D). W programie *Scilab* do wygenerowania charakterystyki częstotliwościowej służy komenda *bode(h),*gdzie w miejsce litery *h* należy wstawić badaną transmitancję.



Rys. 3. Charakterystyka amplitudy i fazy transmitancji $H_g(s)$ przetwornicy BUCK (opis w tekście)

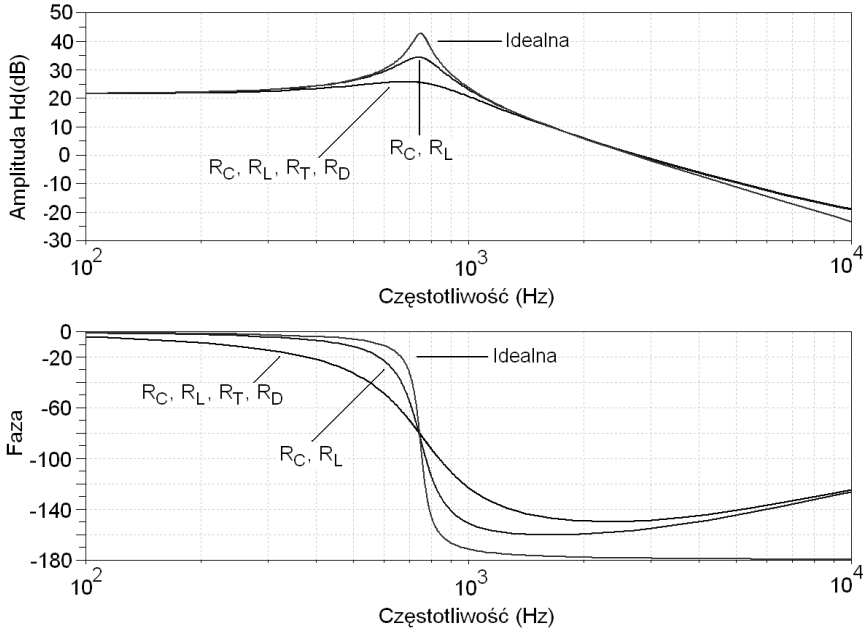
Aby lepiej zrozumieć wykres amplitudy z rys. 3, można posłużyć się przykładem. Na wejście przetwornicy podano napięcie, które posiada składową stałą $V_G=12V$ oraz składową zmienną $v_g(t)$. Składowa zmienna posiada amplitudę 50mV i oscyluje z częstotliwością *f*. Współczynnik wypełnienia d_A jest stały i wynosi 0,5. Jeżeli częstotliwość *f* składowej $v_g(t)$ będzie równa częstotliwości rezonansowej przetwornicy $f=f_r$, to napięcie wyjściowe będzie oscylować z amplitudą:

- ok. 560mV – dla przetwornicy idealnej
- ok. 224mV – dla przetwornicy uwzględniającej tylko niektóre rezystancje pasożytnicze (R_C, R_L)
- ok. 79mV – dla przetwornicy uwzględniającej wszystkie rezystancje pasożytnicze (R_C, R_L, R_T, R_D)

Powyższy przykład wyraźnie pokazuje różnice pomiędzy modelami przetwornicy, opisywanymi w niniejszej pracy. Oczywiście model uwzględniający wyłącznie rezystancję kondensatora i cewki lepiej opisuje działanie przetwornicy, niż model przetwornicy idealnej. Jednak zgodnie z rys. 3 po uwzględnieniu wszystkich rezystancji

pasożytniczych uzyskuje się prawie trzykrotnie mniejszą amplitudę wahań napięcia wyjściowego niż w przypadku uwzględnienia tylko rezystancji kondensatora i cewki.

Analiza częstotliwościowa transmitancji $H_d(s)$ pozwala określić wpływ zmian współczynnika wypełnienia na napięcie wyjściowe przetwornicy w szerokim spektrum częstotliwości. Na rys. 4 przedstawiono wynik symulacji częstotliwościowej transmitancji $H_d(s)$ dla przypadków przetwornicy: idealnej, z rezystancjami pasożytniczymi elementów pasywnych (R_C, R_L) oraz z uwzględnieniem wszystkich rezystancji pasożytniczych (R_C, R_L, R_T, R_D).



Rys. 4. Charakterystyka amplitudy i fazy transmitancji $H_d(s)$ przetwornicy BUCK (opis w tekście)

Analizując charakterystyki z rys. 4 można dojść do podobnych wniosków co w przypadku rys. 3. Według rys. 4 amplituda napięcia wyjściowego dla modelu uwzględniającego wszystkie rezystancje pasożytnicze byłaby ok. 2,5 razy mniejsza niż w przypadku modelu uwzględniającego tylko rezystancję kondensatora i cewki. Oznacza to, że użycie układu sterowania zaprojektowanego na podstawie niepełnego opisu przetwornicy mogłoby powodować większe oscylacje napięcia wyjściowego niż pierwotna przyczyna, tj. wahania współczynnika wypełnienia.

Analiza częstotliwościowa transmitancji $H_d(s)$ oraz $H_g(s)$ pokazuje widoczne różnice pomiędzy transmitancją uwzględniającą jedynie rezystancje pasożytnicze elementów biernych (cewka, kondensator), a transmitancją uwzględniającą wszystkie rezystancje pasożytnicze. Należy zauważyć, że w przypadku uwzględnienia wszystkich rezystancji

pasożytniczych podbicie amplitudy, dla częstotliwości rezonansowej, jest bardzo małe (rys. 3 i 4). W rzeczywistych przetwornicach podbicie amplitudy w charakterystyce częstotliwościowej jest niepożądane ponieważ oznacza, że przetwornica wzmacnia zakłócenia o wyższych częstotliwościach.

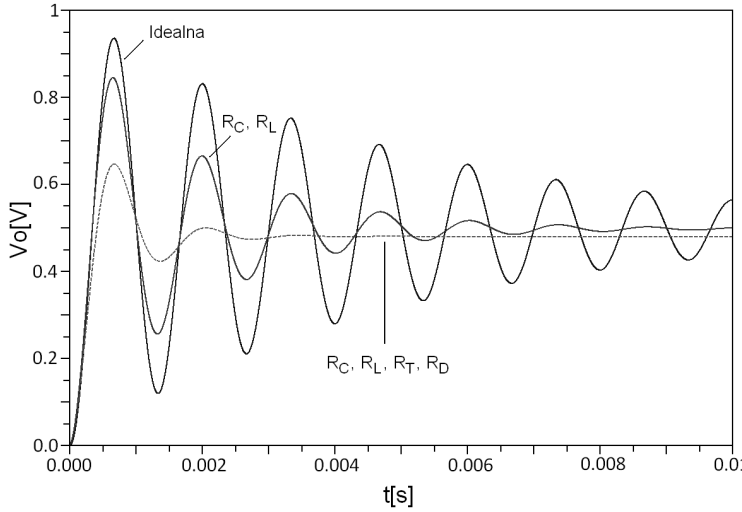
2.3. Symulacje w dziedzinie czasu

W niektórych przypadkach ważne jest aby poznać jak przy pewnych założeniach układ zareaguje np. na skokową zmianę napięcia zasilania. W rzeczywistym obwodzie taka sytuacja może się pojawić, gdy bateria słoneczna bądź turbina wiatrowa nagle zaczną wytwarzać większą moc, czy chociażby przy włączeniu zasilania. Aby uzyskać odpowiedź na to pytanie należy pomnożyć transmitancję $H_g(s)$ przez reprezentację sygnału pobudzającego (uskok jednostkowy) w dziedzinie „s", a następnie wyznaczyć odwrotną transformatę Laplace'a [12](rozdz. 14.5).

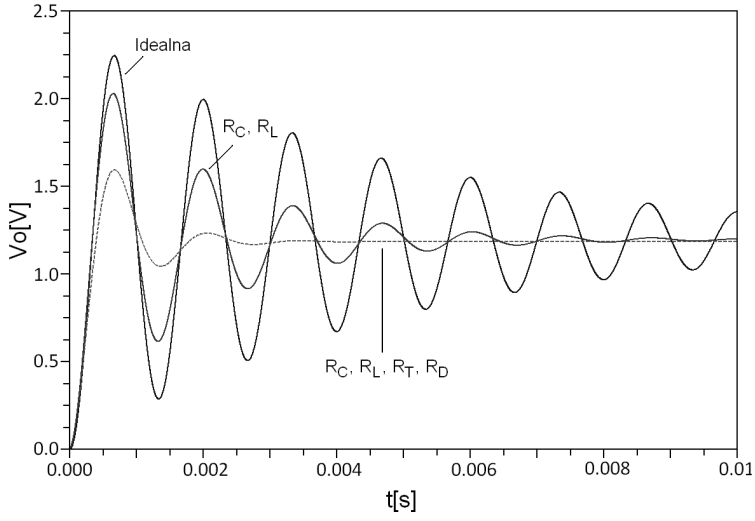
Przy pomocy transmitancji $H_g(s)$ uzyskano odpowiedź napięcia wyjściowego na skokową zmianę napięcia wejściowego o 1V. Wynik symulacji przedstawiono na rys. 5 i dotyczy ona trzech przypadków: przetwornicy idealnej, przetwornicy z uwzględnieniem rezystancji pasożytniczych elementów biernych (R_C, R_L) oraz z uwzględnieniem wszystkich rezystancji pasożytniczych (R_C, R_L, R_T, R_D). W programie *Scilab* do wygenerowania wektora odpowiedzi układu na uskok jednostkowy służy komenda *csim('step',t,h),* gdzie *t* jest wektorem czasu, a *h* oznacza badaną transmitancję.

Na rys. 5 widać znaczną różnicę zarówno w amplitudzie jak i czasie oscylacji napięcia wyjściowego. Różnica w amplitudzie pomiędzy transmitancją uwzględniającą część i transmitancją uwzględniającą wszystkie rezystancje pasożytnicze stanowi ok. 20% amplitudy sygnału pobudzającego co w terminologii miernictwa jest błędem grubym.

Korzystając z transmitancji $H_d(s)$ można uzyskać informację o zachowaniu się napięcia wyjściowego układu po skokowej zmianie współczynnika wypełnienia. Taki przypadek może mieć miejsce, gdy układ sterowania będzie chciał zareagować np. na skokową zmianę obciążenia przetwornicy. Na rys. 6 przedstawiono odpowiedź napięcia wyjściowego na skokową zmianę współczynnika wypełnienia o wartość równą 0,1.



Rys. 5. Odpowiedź napięcia wyjściowego przetwornicy BUCK na skokową zmianę napięcia wejściowego o 1V



Rys. 6. Odpowiedź napięcia wyjściowego przetwornicy BUCK na skokową zmianę współczynnika wypełnienia o wartość 0,1

W przypadku rys. 6 wartość międzyszczytowa wahań napięcia wyjściowego wynosi:

- ok. 2V – dla przetwornicy idealnej
- ok. 1,37V – dla przetwornicy uwzględniającej tylko niektóre rezystancje pasożytnicze (R_C, R_L)
- ok. 0,6V – dla przetwornicy uwzględniającej wszystkie rezystancje pasożytnicze (R_C, R_L, R_T, R_D)

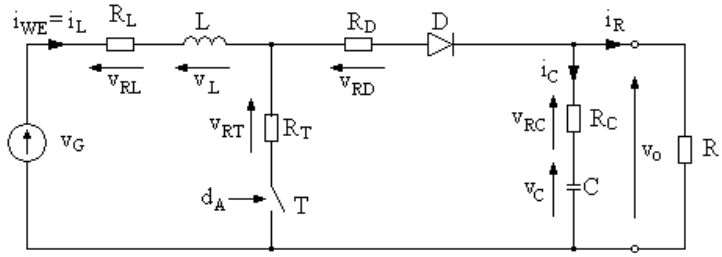
Amplituda wahań napięcia wyjściowego dla przypadku transmitancji uwzględniającej wszystkie rezystancje pasożytnicze jest dwa razy mniejsza niż w przypadku symulacji uwzględniającej jedynie rezystancję kondensatora i cewki. Należy także zauważyć, że również czas trwania oscylacji znacząco różni się pomiędzy wszystkimi wynikami symulacji. W przypadku modelu uwzględniającego wszystkie rezystancje pasożytnicze czas ten jest ponad dwa razy krótszy niż w przypadku modelu uwzględniającego tylko niektóre rezystancje pasożytnicze (R_C, R_L).

W rzeczywistej przetwornicy dąży się do wyeliminowania wszelkich oscylacji, ponieważ mogą one powodować nieprawidłową pracę i wzbudzanie się układów odbiorczych.

3. Przetwornica BOOST

3.1. Postacie transmitancji Hd oraz Hg

Schemat ideowy przetwornicy BOOST, uwzględniający wszystkie rezystancje pasożytnicze, przedstawiono na rys. 7. Parametry elementów przetwornicy zamieszczono w tabeli 2.



Rys. 7. Przetwornica BOOST zawierająca wszystkie rezystancje pasożytnicze

Tabela 2. Parametry elementów przetwornicy BOOST

V_G=3 V	R=5 Ω
D_A=0,5	R_T=28 mΩ
L=50 uH	R_D=300 mΩ
C=487,23 uF	R_L=20 mΩ
	R_C=42,8 mΩ

Transmitancje $H_g(s)$ oraz $H_d(s)$ są definiowane w taki sam sposób jak w przypadku przetwornicy BUCK (1), (3) i dla układu z rys. 7 wynoszą odpowiednio [4]:

$$H_g(s) = \frac{(1-D)(1+sCR_C)}{LC_Z s^2 + (GL + R_Z C_Z + (1-D)^2 CR_C)s + GR_Z + (1-D)^2} \tag{5}$$

$$H_d(s) = \frac{-LCR_C I_L s^2 + (V_A CR_C - LI_L - CR_C R_Z I_L)s + V_A - I_L R_Z}{LC_Z s^2 + (GL + R_Z C_Z + (1-D)^2 CR_C)s + GR_Z + (1-D)^2} \tag{6}$$

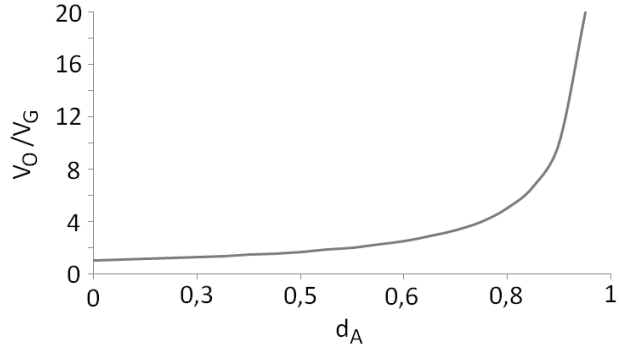
gdzie:

$$V_A = (1-D)(V_O - I_L(R_T - R_D))$$

$$V_O = \frac{(1-D)V_G}{(1-D)^2 + GR_Z}$$

$$I_G = I_L \frac{GV_O}{1-D}$$

W idealnej przetwornicy BOOST zależność wzmocnienia napięcia wejściowego od współczynnika wypełnienia sygnału kluczującego jest nieliniowa (rys. 8). W związku z tym niewielka zmiana współczynnika wypełnienia może powodować duże zmiany napięcia wyjściowego (transmitancja $H_d(s)$).

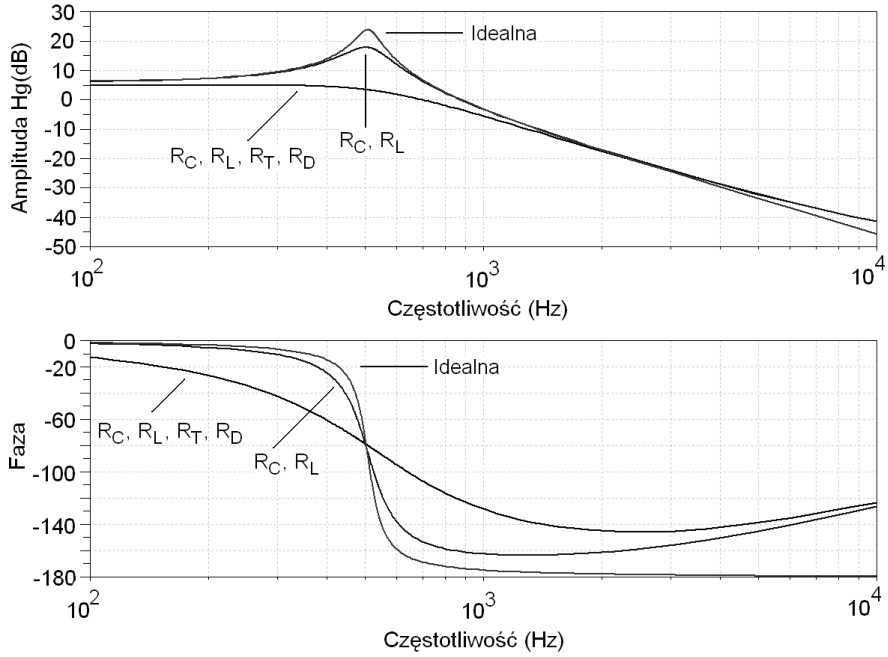


Rys. 8. Zależność wzmocnienia napięcia wejściowego od współczynnika wypełnienia w idealnej przetwornicy BOOST

W rzeczywistym układzie duży współczynnik wypełnienia nie będzie miał takiego wpływu na napięcie wyjściowe jak to przedstawiono na rys. 8. Powodem tego są straty występujące w elementach przetwornicy. Jednak wciąż wpływ współczynnika wypełnienia na napięcie wyjściowe będzie miał charakter nieliniowy i będzie większy niż w przypadku przetwornicy BUCK.

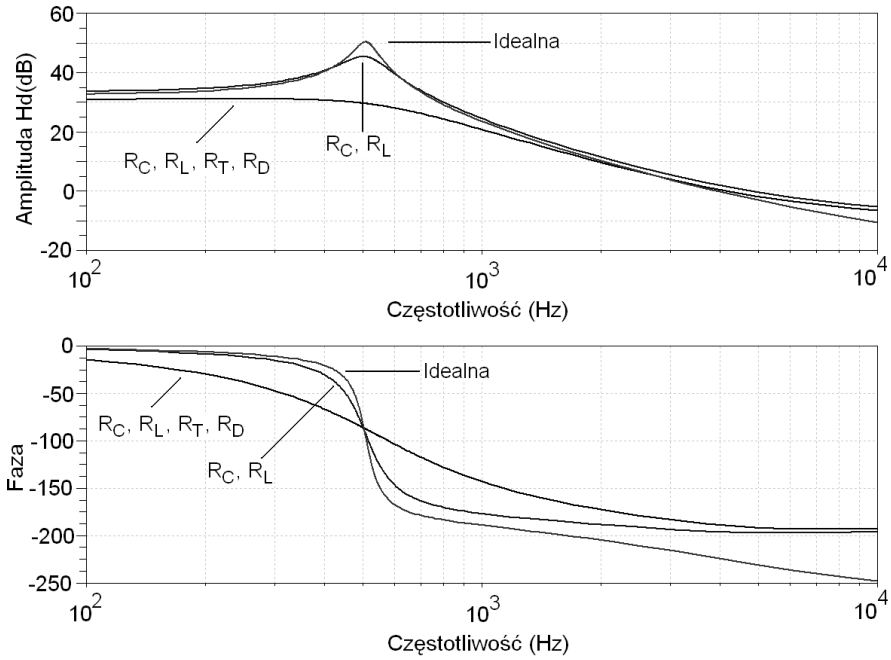
3.2. Symulacje w dziedzinie częstotliwości

Na rys. 9 przedstawiono symulacyjną charakterystykę częstotliwościową transmitancji $H_g(s)$. W symulowanej przetwornicy BOOST indukcyjność cewki została zmniejszona dwukrotnie w stosunku do indukcyjności w przetwornicy BUCK. To powinno spowodować zwiększenie częstotliwości rezonansowej przetwornicy. Tymczasem podbicie amplitudy występuje dla częstotliwości mniejszej w porównaniu z charakterystykami przetwornicy BUCK (rys. 3). W praktyce oznacza to, że przetwornica BOOST jest bardziej podatna na zakłócenia o niskiej częstotliwości niż przetwornica BUCK.



Rys. 9. Charakterystyka amplitudy i fazy transmitancji $Hg(s)$ przetwornicy BOOST (opis w tekście)

Rysunek 10 przedstawia charakterystykę amplitudy i fazy transmitancji $H_d(s)$. W tym przypadku częstotliwość rezonansowa również jest mniejsza niż w przetwornicy BUCK. Niemniej zarówno na rys. 9 jak i na rys. 10 widać znaczną różnicę w kształcie charakterystyk pomiędzy transmitancją uwzględniającą wszystkie rezystancje pasożytnicze (R_C, R_L, R_T, R_D), a pozostałymi transmitancjami, uwzględniającymi tylko część, lub nieuwzględniającymi żadnych rezystancji pasożytniczych. Na obu rysunkach różnica między przypadkiem idealnym, a uwzględniającym wszystkie rezystancje pasożytnicze wynosi ok 20dB, co w przeliczeniu daje 10-krotną różnicę w amplitudzie napięcia.



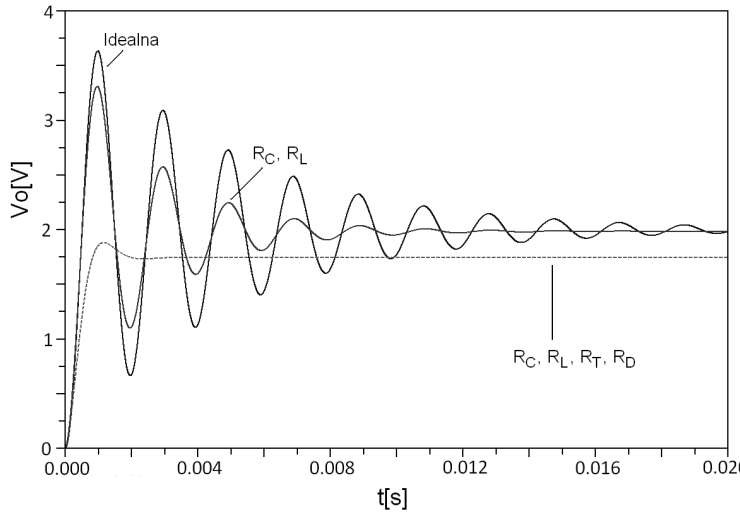
Rys. 10. Charakterystyka amplitudy i fazy transmitancji $H_d(s)$ przetwornicy BOOST (opis w tekście)

Należy również zauważyć, że uwzględnienie wszystkich rezystancji pasożytniczych daje w wyniku zupełnie inny charakter przebiegu amplitudy - zarówno w rys. 9 jak i rys. 10 nie występuje podbicie amplitudy, co jest efektem pożądanym. Dodatkowo na obu rysunkach różnica pomiędzy charakterystykami amplitudowymi uwzględniającymi tylko niektóre oraz wszystkie rezystancje pasożytnicze wynosi ok. 15dB, co w przeliczeniu daje ponad pięciokrotną różnicę w amplitudzie. Niniejszy przykład po raz kolejny pokazuje konieczność uwzględniania rezystancji pasożytniczych w modelach przetwornic.

3.3. Symulacje w dziedzinie czasu

W niniejszym rozdziale wykorzystano transmitancje $H_g(s)$ oraz $H_d(s)$ przetwornicy BOOST do wyznaczenia odpowiedzi napięcia wyjściowego w dziedzinie czasu na skokową zmianę napięcia wejściowego oraz współczynnika wypełnienia.

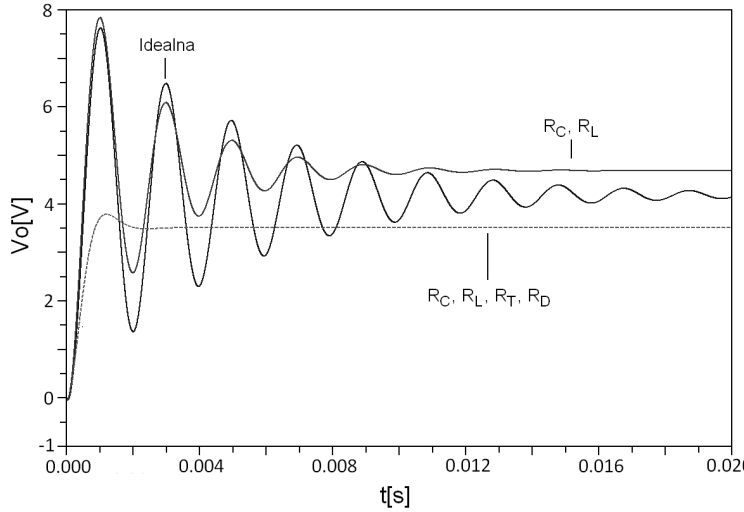
Wpływ skokowej zmiany napięcia wejściowego na napięcie wyjściowe w przetwornicy BOOST przedstawiono na rys. 11. Rysunek ten przedstawia symulację dla trzech przypadków: przetwornicy idealnej, przetwornicy z uwzględnieniem rezystancji pasożytniczych elementów biernych (R_C, R_L) oraz z uwzględnieniem wszystkich rezystancji pasożytniczych (R_C, R_L, R_T, R_D)



Rys. 11. Odpowiedź napięcia wyjściowego przetwornicy BOOST na skokową zmianę napięcia wejściowego o 1V

Z rysunku 11 widać, że uwzględnienie rezystancji kluczy w przetwornicy BOOST w większym stopniu wpływa na kształt jej charakterystyki niż miało to miejsce w przetwornicy BUCK. Dzieje się tak ponieważ w przetwornicy BOOST w fazie przewodzenia łącznika tranzystorowego prąd wejściowy płynie jedynie przez rezystancję cewki oraz rezystancję tranzystora. Jeżeli rezystancje te są tego samego rzędu, to pominięcie którejkolwiek z nich powoduje duże zmiany w wynikach symulacji.

Korzystając z transmitancji $H_d(s)$ uzyskano przebieg czasowy odpowiedzi napięcia wyjściowego przetwornicy BOOST na skokową zmianę współczynnika wypełnienia o wartość 0,1. Wynik symulacji przedstawiono na rys. 12.



Rys. 12. Odpowiedź napięcia wyjściowego przetwornicy BOOST na skokową zmianę współczynnika wypełnienia o 0,1

W przetwornicy BOOST jakiekolwiek zmiany współczynnika wypełnienia bardziej wpływają na napięcie wyjściowe niż ma to miejsce w przetwornicy BUCK. Wynika to z faktu, że w przetwornicy BOOST zależność wzmocnienia napięcia wejściowego od współczynnika wypełnienia jest nieliniowa, co zilustrowano na rys. 8 dla przypadku idealnego. Dlatego niewielkie zmiany współczynnika wypełnienia w pewnym obszarze skutkują dużymi zmianami napięcia wyjściowego.

4. Podsumowanie

Przedstawione w powyższych rozdziałach symulacje dotyczą przetwornic BUCK i BOOST zawierających rezystancje pasożytnicze wszystkich elementów - zarówno aktywnych jak i pasywnych. Otrzymane charakterystyki dowodzą, że uwzględnianie jedynie części rezystancji pasożytniczych powoduje znaczną różnicę w odniesieniu do przebiegów uwzględniających wszystkie rezystancje pasożytnicze. Opisano znaczenie symulacji w dziedzinie częstotliwości i w dziedzinie czasu. Jednym z kolejnych etapów pracy będzie zbadanie wpływu poszczególnych rezystancji pasożytniczych na częstotliwość rezonansową przetwornicy oraz zbadanie wpływu rezystancji pasożytniczych na pracę przetwornicy znajdującej się w trybie nieciągłego prądu cewki DCM.

Bibliografia

1. Erickson R.W., Maksimovic D., *Fundamentals of Power Electronics Second Edition*, University of Colorado, Boulder
2. Biolkova V., Kolka Z., Biolek D., *State-Space Averaging (SSA) Revisited: On the Accuracy of SSA-Based Line-To-Output Frequency Responses of Switched DC-DC Converters*, WSEAS TRANSACTIONS on CIRCUITS and SYSTEMS , Issue 2, VOL. 9, February 2010
3. Janke W., *Averaged models of pulse-modulated DC-DC power converters. Part I. Discussion of standard methods*, Archives Of Electrical Engineering VOL 61 (4), pp. 609-631 (2012)
4. Janke W., *Averaged models of pulse-modulated DC-DC power converters. Part II: Models based on the separation of variables*, Archives Of Electrical Engineering VOL. 61 (4), pp. 633-654 (2012)
5. Middlebrook R.D., Ćuk S., *A General Unified Apprroach To Modelling Switching-Converter Power Stages*, IEEE Power Electronics Specialists Conference, June 8-10, 1976, Cleveland, OH.
6. Vorperian V., *Simplified Analysis of PWM Converters Using Model of PWM Switch Part I: Continuous Conduction Mode*, IEEE Transactions On Aerospace And Electronic Systems VOL. 24, N0. 3 May 1990
7. Dijk E., Spruijt J.N., Sullivan M.O., *PWM-switch modeling of DC-DC converters*, IEEE Transactions On Power Electronics, VOL. 10, NO. 6, November 1995
8. Jinno M., Chen P., Lai Y., Harada K., *Investigation on the Ripple Voltage and the Stability of SR Buck Converters Witch High Output Current and Low Output Voltage*, IEEE Transactions On Industrial Electronics. VOL. 57. NO3. March 2010
9. Janke W., Bączek M., Walczak M., *Output characteristics of step-down (Buck) power converter*,
10. Janke W., Walczak M., Bączek M., *Wpływ elementów pasożytniczych na charakterystyki wejściowe i wyjściowe impulsowych przetwornic napięcia*, Modelowanie, Symulacja I Zastosowania W Technice, Kościelisko, 18-22 czerwca 2012r.
11. Walczak M., *Symulacje charakterystyk wejściowych i wyjściowych impulsowych przetwornic napięcia stałego w trybie CCM*, XIV International PhD Workshop OWD 2012, 20-23 October 2012
12. Bolkowski S., *Teoria Obwodów Elektrycznych*, wyd. ósme, Warszawa, WNT. 2005

Abstract

Two popular DC-DC power converters (BUCK and BOOST) have been discussed. Models of line-to-output and control-to-output transfer functions, considering all parasitic resistances, have been recalled. Based on those models examples of simulations corresponding to characteristics in frequency and time domain have been presented. The simulations have been performed with free of charge program called *Scilab*. All parameters of converters' elements used for the simulations, including parasitic resistances, have been acquired during measurements. It means that simulated converters refer to a good approximation of real device which has been built and tested. Practical meaning of every simulation has been explained.

The paper reveals differences between simulations of models considering three types of mathematical models which include: all, some of and none of a parasitic resistances. It is shown that simulations generated according to models considering only parasitic resistances of inductor and capacitor are substantially different from simulations achieved based on model, which additionally considers parasitic resistances of switching elements (transistor and diode).

Key words: BUCK, BOOST, Step-down, Step-up, DC-DC converter, characteristics of converters, parasitic resistance.

Streszczenie

W niniejszej pracy przedstawiono transmitancje opisujące wpływ zmian napięcia wejściowego oraz współczynnika wypełnienia na amplitudę napięcia wyjściowego dwóch popularnych przetwornic BUCK i BOOST. Wspomniane transmitancje są modelami małosygnałowymi przebiegów uśrednionych na okres kluczowania wyżej wymienionych przetwornic. Przy pomocy programu *Scilab* wykonano symulacje tych modeli w dziedzinie częstotliwości i czasu. Przedstawione wyniki symulacji stanowią odpowiedź układu na skokowe oraz okresowe zmiany zarówno napięcia wejściowego jak i współczynnika wypełnienia oraz przedstawiają różnice pomiędzy trzema modelami przetwornic, które uwzględniają: wszystkie, część lub nie uwzględniają żadnych rezystancji pasożytniczych. Znaczenie poszczególnych charakterystyk przetwornic zostało wyjaśnione w tekście pracy.

Słowa kluczowe: BUCK, BOOST, przetwornica, Step-down, Step-up, charakterystyki przetwornic, rezystancje pasożytnicze.

Michał Pawlak
Wydział Fizyki, Astronomii i Informatyki Stosowanej
Uniwersytet Mikołaja Kopernika
ul. Grudziądzka 5
87-100 Toruń
Polska

Mirosław Andrzej Maliński
Wydział Elektroniki i Informatyki
Politechnika Koszalińska
ul. Śniadeckich 2
75-343 Koszalin
Polska

Wyznaczanie termodyfuzyjności kryształów CdMgSe z wykorzystaniem techniki radiometrii w podczerwieni

Słowa kluczowe: termodyfuzyjność, kryształy półprzewodnikowe, CdMgSe, radiometria w podczerwieni.

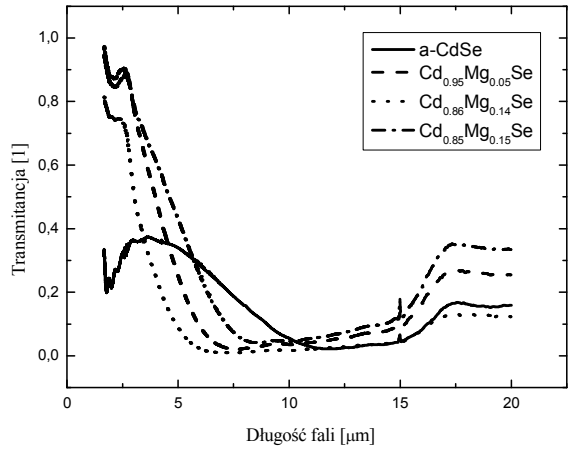
Wstęp

Termodyfuzyjność D_t jest wielkością fizyczną opisującą właściwości termiczne materiałów elektronicznych. Termodyfuzyjność może zostać wyznaczona za pomocą technik eksperymentalnych opartych na metodzie fal termicznych. Ich zaletą jest nieniszczący charakter badania. Spośród dostępnych technik największe znaczenie ma technika fotopiroelektryczna PPE (ang. photopyroelectric) [1]. Pomiar fali termicznej w tej metodzie polega na rejestracji sygnału elektrycznego powstałego w wyniku zmian temperatury detektora piroelektrycznego. Zmiany temperatury detektora piroelektrycznego spowodowane są periodycznym oświetlaniem próbki. Technika ta wymaga kontaktu pomiędzy badanym materiałem oraz detektorem. W celach praktycznych znacznie bardziej pożądane jest aby pomiar odbywał się w sposób bezkontaktowy. Spośród dostępnych technik radiometria w podczerwieni PTR (ang. photothermal radiometry) spełnia to wymaganie [2]. Pomiar fali termicznej w tej metodzie polega na rejestracji promieniowania podczerwonego emitowanego z próbki. W porównaniu z sygnałem PPE, sygnał PTR badanej próbki zależy od jej współczynnika absorpcji w obszarze pracy detektora podczerwieni [3].

Obiekt badań oraz jego właściwości

Kryształy półprzewodnikowe CdSe oraz $Cd_{1-x}Mg_xSe$ zostały otrzymane w Zakładzie Fizyki Półprzewodników i Fizyki Węgla Instytutu Fizyki UMK zmodyfikowaną metodą Bridgmana [4] w układzie pionowym. Następnie kryształy były pocięte na płytki o grubości około 1-1.5 mm każda. Ucięte próbki szlifowano proszkiem Al_2O_3 o średnicy 10 μm, natomiast aby uzyskać powierzchnię zwierciadlaną polerowano je mechanicznie pastą diamentową (średnica ziaren między 0.1 a 1 μm). Wypolerowane próbki poddano dalszej obróbce chemicznej, która polegała na wytrawieniu ich w mieszaninie $K_2Cr_2O_7$, H_2SO_4 oraz H_2O (w stosunku 3:2:1) zwanej dalej chromianką, oraz 50 % roztworze NaOH. Po trawieniu próbki myto w wodzie destylowanej i alkoholu.

Właściwości optyczne w obszarze podczerwonym kryształów półprzewodnikowych CdSe oraz $Cd_{1-x}Mg_xSe$ zostały zbadane za pomocą spektroskopii fourierowskiej. Widma spektralne zostały otrzymane za pomocą IRScope II umieszczonego w spektrometrze IFS 66 (Bruker GmbH, Ettlingen, Niemcy) wyposażonym w detektor MCT chłodzony ciekłym azotem. Widma była zarejestrowane w modzie transmisyjnym. Na rysunku 1 zostały przedstawione widma transmisyjne kryształów CdMgSe.



Rys. 1. Widma transmisji kryształów CdMgSe

Współczynnik absorpcji może zostać obliczony za pomocą metody zaproponowanej przez Malińskiego w pracach [5, 6]. Uzyskany tą metodą współczynnik absorpcji wynosi od 5 cm^{-1} do 40 cm^{-1} w zależności od kryształu oraz długości fali. Oznacza to, że kryształy CdMgSe w obszarze podczerwonym są półprzepuszczalne dla promieniowania podczerwonego.

Teoria

Analiza sygnału PTR emitowanego przez próbki przeźroczyste lub półprzeźroczyste dla promieniowania podczerwonego jest bardzo trudna. W celu uproszczenia analizy teoretycznej do badanych próbek zostały przytwierdzone, za pomocą specjalnego smaru APIEZON, cienkie warstwy folii aluminiowej (Rysunek 2).

Rys. 2. Trójwarstwowy model fizyczny

Przy częstotliwościach modulacji mniejszych od 10 Hz model fizyczny próbki z dwiema cienkimi (grubość foli wynosiła 25 μm) warstwami folii aluminiowej o dobrym przewodnictwie cieplnym (układ trójwarstwowy), przedstawiony na rysunku 2, może być zredukowany do układu jednowarstwowego. W takim przypadku fala termiczna może być opisana przez wyrażenia podane przez Malińskiego [7]

$$\Delta T(z,f) = \frac{\beta \cdot I \cdot (M+N)}{2 \cdot \lambda \cdot \sigma \cdot (1 - 2 \cdot \exp(-2 \cdot \sigma \cdot L))} \tag{1}$$

$$M = \frac{\exp(-\sigma z) - \exp(-z\beta) + \exp\left[-\sigma(2L-z)\right] - \exp(-\sigma L + \sigma z - L\beta) + \exp(-2\sigma L - z\beta)}{\beta - \sigma} \\ \frac{-\exp(-\sigma L - \sigma z - L\beta)}{\beta - \sigma} \tag{1a}$$

$$N = \frac{\exp(-\sigma z) + \exp(-2\sigma L + \sigma z) - \exp(-2\sigma L - z\beta) + \exp(-\beta z) - \exp\left[-(\sigma+\beta)L + \sigma z\right]}{\beta + \sigma} \\ \frac{-\exp(-\sigma z - \sigma L - L\beta)}{\beta + \sigma} \tag{1b}$$

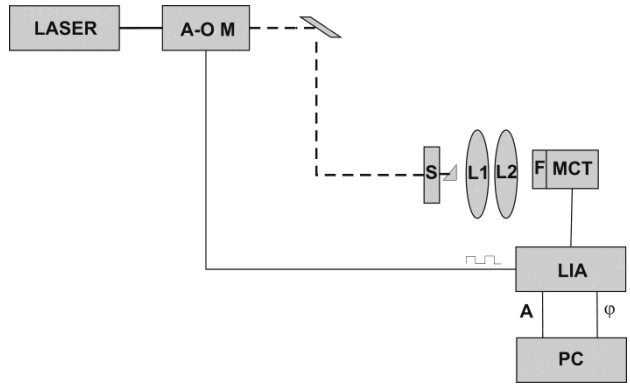
$$\sigma(f) = \sqrt{\frac{D_t}{\pi f}} \tag{1c}$$

gdzie β jest współczynnikiem absorpcji dla światła wzbudzającego, λ – przewodnością cieplną, L – grubością próbki. Dla próbki optycznie nieprzeźroczystej (dla światła wzbudzającego) oraz termicznie grubej faza fali termicznej φ opisana wzorem (1) można napisać:

$$\phi = -\frac{\pi}{2} - \sqrt{\frac{\pi f}{D_t}} L_s = -\frac{\pi}{2} - m\sqrt{f} \tag{2}$$

Układ eksperymentalny

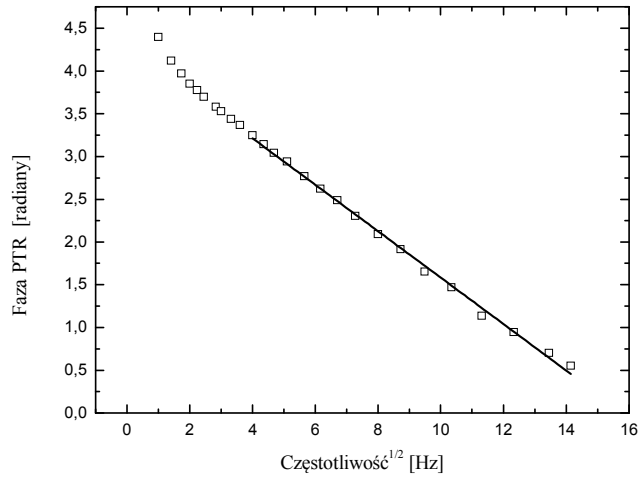
Schemat układu pomiarowego został przedstawiony na rysunku 3.



Rys. 3. Schemat układu pomiarowego w konfiguracji transmisyjnej

Fala termiczna jest wzbudzana w próbce (*S*) za pomocą wiązki lasera argonowego (moc 200 mW oraz średnica wiązki około 2 mm). Wiązka laserowa jest intensywnie modulowana za pomocą modulatora akustooptycznego (*AOM*). Następnie wiązka jest kierowana na próbkę za pomocą pryzmatu. Sygnał PTR jest emitowany przez drugą powierzchnię oraz zbierany za pomocą dwóch soczewek wykonanych z BaF_2. Następnie sygnał skupiany był na detektorze MCT pracującym w zakresie od 2 μm do 12 μm oraz jest analizowany za pomocą wzmacniacza fazoczułego (*LIA*, ang. lock in-amplifier) oraz komputera (*PC*).

Na rysunku 4 przedstawiona została faza sygnału PTR wraz z najlepszym dopasowaniem liniowym metodą najmniejszych kwadratów do danych eksperymentalnych dla próbki molibdenu o grubości L_s=0.11 cm.



Rys. 4. Faza sygnału PTR próbki molibdenu o grubości L_s=0.11 cm

W wyniku najlepszego dopasowania liniowego metodą najmniejszych kwadratów do funkcji y(x)=m·x + b uzyskano: m=0.27±0.01 oraz b = 4.3 ±0.3.

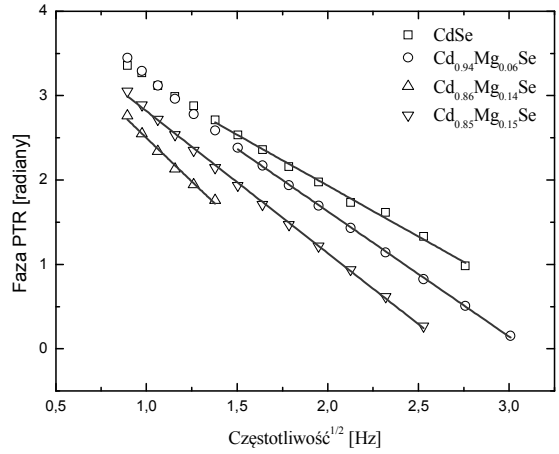
Termodyfuzyjność dla próbki molibdenu może być wyznaczona za pomocą wzoru:

$$D_t = \frac{\pi \cdot L_s^2}{m^2} \tag{3}$$

gdzie *m* jest współczynnikiem kierunkowym prostej zależności fazy sygnału PTR od pierwiastka z częstotliwości zgodnie z (2). Termodyfuzyjność wyznaczona za pomocą wzoru (3) wynosi $5.2 \cdot 10^{-5}$ m^2/s i jest zgodna z danymi dostępnymi w literaturze. Otrzymany wynik potwierdza skuteczność techniki PTR w wyznaczaniu termodyfuzyjności próbek nieprzeźroczystych w obszarze podczerwieni.

Wyniki eksperymentalne oraz dyskusja

Na rysunku 5 zostały przedstawione fazy sygnału PTR zarejestrowane dla kryształów CdMgSe z foliami aluminiowymi na ich powierzchniach wraz z najlepszym dopasowaniem liniowym metodą najmniejszych kwadratów do (2).



Rys. 5. Fazy sygnału PTR zarejestrowane dla kryształów CdMgSe w konfiguracji transmisyjnej wraz z najlepszym dopasowaniem liniowym do (2)

Wartości parametru *m* oraz termodyfuzyjności wyznaczonej za pomocą wzoru (3) zostały zamieszczone w tabeli 1.

Tab. 1. Termodyfuzyjność kryształów CdMgSe wyznaczona z fazy sygnału PTR zarejestrowanego w konfiguracji transmisyjnej

Próbka	*m*	D_{PTR} ($cm^2 \cdot s^{-1}$)
CdSe	-1.20	0.038
$Cd_{0.94}Mg_{0.06}Se$	-1.48	0.016
$Cd_{0.86}Mg_{0.14}Se$	-2.06	0.013
$Cd_{0.85}Mg_{0.15}Se$	-1.68	0.010

W tabeli 2 zostały przedstawione wyniki uzyskane za pomocą techniki fotopiroelektrycznej (PPE) [8] oraz techniki PTR w konfiguracji transmisyjnej.

Tab. 2. Porównanie wartości termodyfuzyjności otrzymanych za pomocą radiometrii w podczerwieni w konfiguracji transmisyjnej oraz techniki fotopiroelektrycznej (PPE)

Próbka	D_{PPE} ($cm^2 \cdot s^{-1}$)	D_{PTR} ($cm^2 \cdot s^{-1}$)
CdSe	0.047	0.038
$Cd_{0.94}Mg_{0.06}Se$	0.022	0.016
$Cd_{0.86}Mg_{0.14}Se$	0.017	0.013
$Cd_{0.85}Mg_{0.15}Se$	0.016	0.010

Otrzymane wartości termodyfuzyjności dla kryształów półprzewodnikowych CdMgSe za pomocą technik PPE i PTR różnią się w niewielkim stopniu. Należy jednak podkreślić, że do analizy sygnału PTR został wykorzystany model jednowarstwowy oraz analizie podlegała jedynie faza sygnału PTR. W rzeczywistości do analizy sygnału PTR w konfiguracji transmisyjnej znacznie lepszy byłby model trójwarstwowy, np. opracowany przez Malińskiego [9], oraz analizie powinna podlegać zarówno amplituda jak i faza sygnału PTR.

Podsumowanie

W pracy została wyznaczona termodyfuzyjności kryształów półprzewodnikowych w sposób bezkontaktowy za pomocą radiometrii w podczerwieni w konfiguracji transmisyjnej. W celu minimalizacji efektu związanego z półprzeźroczystością promieniowania podczerwonego do powierzchni kryształów zostały przyklejone folie aluminiowe. Do analizy fazy sygnału PTR został wykorzystany model jednowarstwowy. Uzyskane wyniki termodyfuzyjności w niewielkim stopniu różnią się od wartości otrzymanych za pomocą techniki fotopiroelektrycznej.

Literatura

1. Mandelis, M. Zver, “Theory of photopyroelectric spectroscopy of solids”, J. Appl. Phys. 57 (1985), 4421-4431.
2. P. E. Nordal, S. O. Kanstad, “Photothermal Radiometry“, Phys. Scr. 20 (1979), 659-662.
3. R. Fuente, A. Mendioroz, E. Apiñaniz, A. Salazar, “Simultaneous Measurement of Thermal Diffusivity and Optical Absorption Coefficient of Solids Using PTR and PPE: A Comparison”, Int. J. Thermophys. 33 (2012), 1876–1886.
4. F. Firszt, A. Wronkowska, A. Wronkowski, Łęgowski S., Marasek A., H. Męczyńska, M. Pawlak, W. Paszkowicz, J. Zakrzewski, K. Strzałkowski, “Growth and Optical Characterization of CdBeSe and CdMgSe crystals”, Cryst. Res. Technol. 40 (2005), 386-394.

5. M. Maliński, Ł. Chrobak "Transmission and Absorption Based Photoacoustic Methods of Determination of the Optical Absorption Spectra of Si samples-comparison", Solid State Communications 149 (2009), 1600 – 1604.
6. M. Maliński, Ł. Chrobak, J. Zakrzewski, K. Strzałkowski "Determination of the Quantum Efficiency of Luminescence in Mn2+ Ions in ZnBeMnSe Crystals by the Nondestructive Photoacoustic Method", Optical Materials 33 (2010), 75-78.
7. M. Maliński, "Temperature distribution formulae-applications in photoacoustics", Archives of Acoustics 27 (2002), 217-228.
8. M. Pawlak, F. Firszt, S. Łęgowski, H. Męczyńska, J. Gibkes, J. Pelzl, "Thermal Transport Properties of CdMgSe Mixed Crystals Measured by Means of the Photopyroelectric Method", Int. J. Thermophys. 31 (2010),187-198.
9. M. Maliński, "Fotoakustyka i spektroskopia fotoakustyczna materiałów półprzewodnikowych", WPK, Koszalin 2004.

Streszczenie

W pracy została wyznaczona termodyfuzyjność mieszanych kryształów półprzewodnikowych CdMgSe za pomocą radiometrii w podczerwieni. Badane próbki w obszarze widmowym pracy detektora są półprzepuszczalne dla promieniowania podczerwonego. Otrzymane wartości termodyfuzjności dla kryształów półprzewodnikowych CdMgSe za pomocą technik PPE i PTR różnią się w niewielkim stopniu.

Abstract

In this paper the thermal diffusivities of CdMgSe mixed crystals were determined by means of the photothermal radiometry. Investigated samples are semi-transparent in the infrared range. Obtained results are in a reasonable agreement with those determined from the photopyroelectric technique (PPE).

Keywords: thermaldiffusivity, semiconductor crystals, CdMgSe, IR radiometry

Łukasz Przeniosło
Marcin Walków
Sonia Krzeszewska
Sandra Pisarek
Biomedical Engineering Student Research Group "AKSON"
Andrzej Biedka
Marek Jaskuła
Daniel Matias
Krzysztof Penkala
Department of Systems, Signals and Electronics Engineering
Faculty of Electrical Engineering,
West Pomeranian University of Technology,
37 Sikorskiego Str.
70-313 Szczecin

Integrated impedance scanner in selected biomeasurement applications – control circuit

Keywords: bioelectrical impedance, impedance spectroscopy, tissue, analysis of body composition, BIA - Bioelectrical Impedance Analysis, impedance scanner, microprocessor system, SoC, simulation.

1. Introduction

Due to its biological structure and electrical properties, living tissue can be characterized by the Bioelectrical Impedance (BI). Bioimpedance measurements are used in medicine for diagnostic purposes, including the so-called Bioelectrical Impedance Analysis (BIA), which is a noninvasive method of body composition analysis [1-3]. This method is increasingly popular also in sports medicine, aesthetics medicine and dietetics. Several manufacturers offer specialized electronic equipment for professional body composition measurements based on bioimpedance analysis [3]. However, the use of such popular, cheap systems is rather restricted to the dedicated area, and – as a result – in more universal studies one has to adapt the equipment to the particular purpose, which could be difficult, or to purchase much more sophisticated system at much higher price.

The aim of the presented study is to determine and discuss the capabilities of AD5933 integrated circuit (Analog Devices), a System-on-a-Chip (SoC) impedance scanner, in selected Bioelectrical Impedance measurement tasks concerning the above mentioned fields, with a perspective of building a cheap and universal BIA system. The

design of an electronic circuit based on ARM microprocessor for control of AD5933 operation is also presented in the paper.

2. Material and methods

In the study, a simplified model of the human body was used [4], namely a two-terminal RC circuit (Figure 1). In the impedance measurements, range of frequency from 10 kHz to 90 kHz was applied. However, as results from the literature, in body composition analysis (BIA), the optimal frequency is equal to 50 kHz [2,3,5].

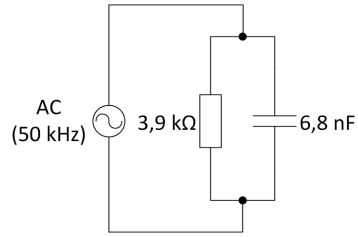


Fig. 1. Electrical model of the human body

Comparative measurements of impedance for the electrical model of tissue were performed with the use of integrated circuit AD5933, placed on the evaluation board EVAL-AD5933EBZ (Analog Devices), and Bodycomp-MF, MF-type (MultiFrequency: 5, 50 and 100 kHz) instrument manufactured by the AKERN company, Italy [2,5]. The device is dedicated to bioimpedance measurements of the body, and equipped with the BodyGram software for processing of the results, mainly calculations of body composition.

Similar comparative measurements were carried out for standard samples - physiological solutions (saline solutions with molar concentrations equal to 0,002 M and 0,01 M) in the laboratories of the Department of Medical Physics and Biophysics of the Pomeranian Medical University in Szczecin. In this study, additionally the Quantum II from JBL company (USA) was used. This is the SF-type BIA instrument (SingleFrequency: 50 kHz) [5]. For comparison, only 50 kHz was used.

3. Results

The results of impedance measurements for electrical model of the tissue (Figure 1), performed with the use of integrated circuit AD5933, were compared with the results of simulation carried out in the PSpice program for this model. Results of this comparison are shown on Figure 2 and Figure [5]. Both impedance frequency characteristics showed satisfactory similarity.

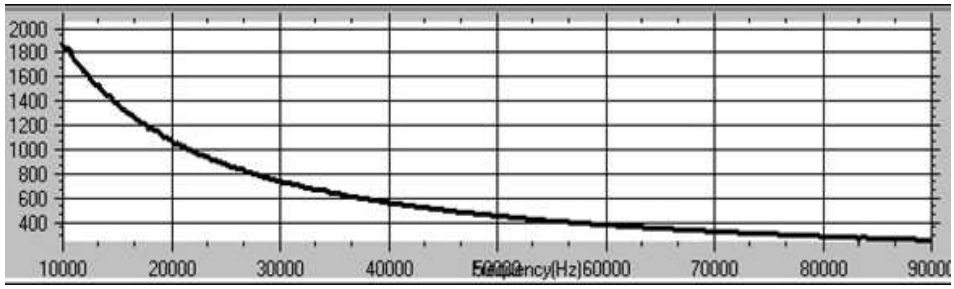


Fig. 2. Measurements performed with the AD5933 integrated circuit

Similar compatibility of the results was obtained in comparative measurements with the use of AD5933, Bodycomp-MF and Quantum II for physiological solutions (only single frequency: 50 kHz). Maximum differences of the values of impedance did not exceed a few percent [5].

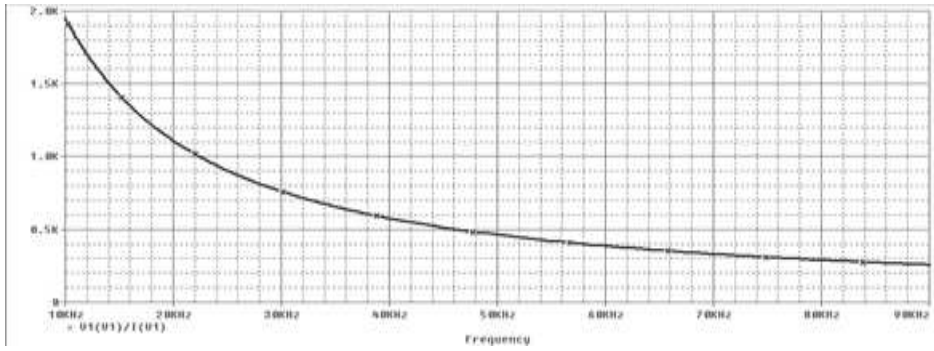


Fig. 3. Results of simulation in the PSpice program

4. Control circuit

The design of an electronic circuit based on the ARM microprocessor for control of the AD5933 operation was another goal of this research work. First, some important assumptions regarding the project of control circuit had to be made.

The main requirement met by the microcontroller for this project, is to have the hardware implemented I2C bus, or any other compliant bus. This lets us write the software faster, as we don't have to bother with software implementation of I2C. It is needed to communicate with AD5933 module, to take measurements.

Number of pins belonging to the MCU has to be at least sufficient - that is, they have to cover LCD graphical display communication and couple I/O's for tact switch keyboard to control the device.

Hardware implemented SPI bus frees from writing a software version of it. This bus is necessary to write and read measurements data from SD card for later use at a PC.

From two to three timers are needed to properly assign numerous tasks for the MCU. Pulse width modulation (PWM) is needed to control the LCD graphical display backlit brightness, and one of the timers can be used in PWM mode.

Assuming the user would need to upgrade his devices software, sending it to the manufacturer would leave user without the equipment for a longer period of time. That's why it is important to have the ability to update devices software without contribution of the manufacturer. That can be achieved, by having device USB module on board. MCU can be connected to the PC via USB and with bootloader on the microcontroller side, and dedicated application on the PC side, software update can be done.

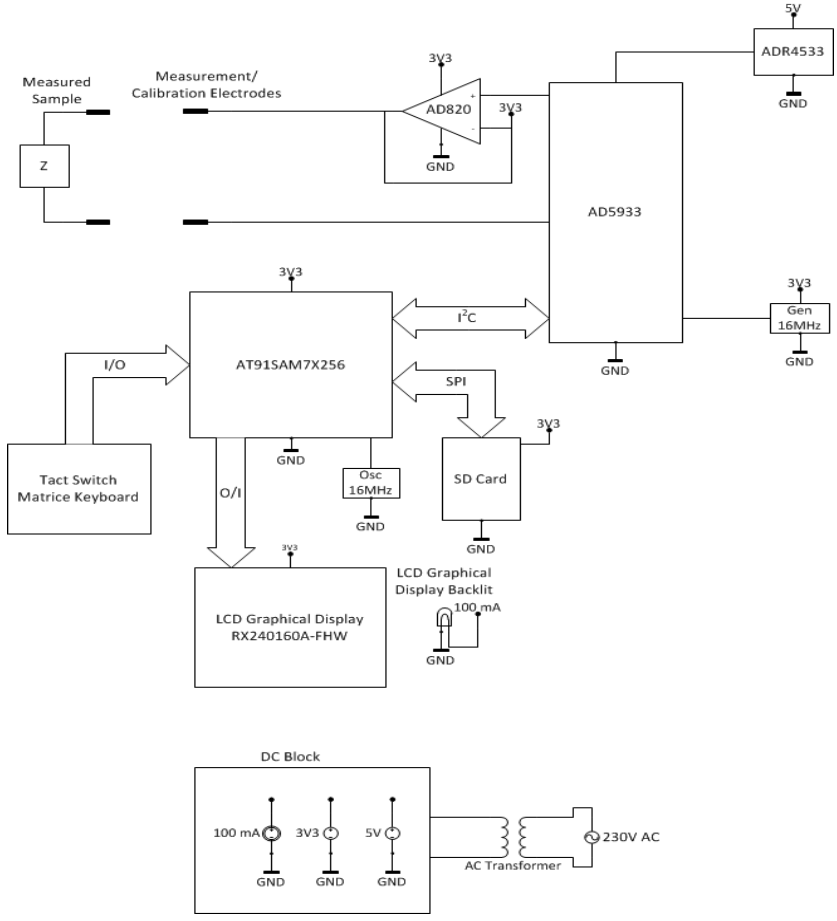


Fig. 4. Block diagram of the microprocessor-based BI-measurement system (explanations in the text)

This application at some point can be quite computationally extensive (i.e. drawing plots on the display). Also, it has to do some arithmetic operations. In that case, 32-bit microcontroller is preferable.

At last, MCU's memory has to be big enough to store the compiled program, and be able to execute all assumed tasks within least external components.

Taking under consideration all of the requirements met by this project, we decided to use for our prototype (and preferably for the final stationary version of the instrument) the AT91SAM7256 microcontroller produced by Atmel (ARM architecture). This module covers all needs of the application, leaving some "free space" for hardware and software expansions which are likely to take place when using the device for research purposes. Also its capacities are much better than AVR and PIC microcontrollers for the same price.

On Figure 4 the schematic diagram of the microprocessor BI-measurement system is shown.

For development work on the project, two evaluation boards were used: EVAL-AD5933EBZ (Analog Devices) with the integrated impedance scanner AD5933 and a popular development board with the AT91SAM7256 microcontroller (Atmel). Additional devices and modules shown on Figure 4 are described below.

ADR4533 – reference voltage (3.3 V); proper measurement can be done only if stable (not changing over time) voltage is applied to the analog and digital sections of the system; AD5933 operates with very low current, that is why it can be also supplied with the ADR4533.

Generator 16 MHz – system clock for the AD5933 circuit.

AD820 – used as an amplifier/buffer for the output excitation signals generated by the AD5933 scanner.

5. Conclusions

The results of the comparative impedance measurements fully justify the use of AD5933 module in BIA research. System-On-a-Chip (SoC), implemented in the layout, is a full-featured impedance spectroscopy measuring device.

Research tasks for further work, in which it is planned to use the results of current studies, include:

- construction of the BIA device for assessment of body composition; this application is important in the field of medical diagnostics, dietetics as well as aesthetic medicine;
- development of measuring instrument for assessment of fatigue degree after exercise (physical effort); as part of a separate study, using adapted Bodycomp-MF apparatus, we obtained very promising results for measuring with BI

technique the lactate and anaerobic thresholds of blood (LT and AT); results of this research are important mainly in sports medicine;

- development of a system for BI measurement using graphene-based biosensor; the BI-SENSOR project is currently run at the West Pomeranian University of Technology, Szczecin, under the GRAPH-TECH research and development program; such application is extremely important in advanced medical diagnostics.

In all the above specified tasks, two versions of equipment will be designed on the basis of AD5933 scanner: universal systems for stationary laboratory use, and mobile instruments for simple, fast field tests (mainly PoC, i.e. Point-of-Care devices).

Thanks to innovative solutions, it will be possible to achieve commercialization effects for the results of presented research.

References

1. Kyle U.G. et al.: Bioelectrical Impedance Analysis, part I: Review of Principles and Methods. Elsevier Ltd, 2004.
2. Lewitt A., Mądro E., Krupienicz A.: Podstawy teoretyczne i zastosowania analizy impedancji bioelektrycznej (BIA).Via Medica, Warszawa, 2007.
3. Lewandowski D., Biedka A., Borkowska K., Penkala K.: Testy funkcjonalne dwóch wybranych urządzeń do analizy impedancji bioelektrycznej (BIA). Inżynieria Biomedyczna - Acta Bio-Optica et Informatica Medica, vol. 16, No. 2, 2010, pp. 83-87.
4. Pilawski A. (ed.): Podstawy Biofizyki. PZWL, Warszawa, 1977.
5. Walków M., Przeniosło Ł., Penkala K.: Badanie możliwości funkcjonalnych scalonego skanera impedancji w pomiarach wybranych charakterystyk organizmu człowieka. Majówka Młodych Biomechaników, Ustroń, 9-13 maja 2013. In press.

The work was in part supported by the grant “Multifunctional graphene-based biosensor for medical diagnosis (BI-SENSOR)” from the National Centre for Research and Development (NCBiR)

Abstract

In the paper the results of preliminary experiments concerning bioelectrical impedance measurements using integrated impedance scanner AD5933 (Analog Devices) are described. Results of performed simulations are also presented. Design of a control circuit based on ARM microprocessor is described. Possible applications of the bioelectrical impedance spectroscopy method in measurements of selected characteristics of the human organism are discussed as well as plans for future development of dedicated, specialized equipment.

Streszczenie

W pracy przedstawiono wyniki wstępnych eksperymentów dotyczących pomiarów impedancji bioelektrycznej przy użyciu zintegrowanego skanera AD5933. Zaprezentowano wyniki przeprowadzonych symulacji oraz konstrukcję układu sterowania w oparciu o mikroprocesor. Omówiono możliwe zastosowania tej metody spektroskopii impedancji bioelektrycznej w pomiarach wybranych cech ludzkiego organizmu, a także możliwości przyszłego rozwoju dedykowanego urządzenia specjalistycznego.

Słowa kluczowe: impedancja bioelektryczna, spektroskopia impedancji, tkanki, analiza składu ciała, BIA – analiza impedancji bioelektrycznej, skaner impedancyjny, system mikroprocesorowy, symulacje

Sylwester Wosiak
Wydział Informatyki
Zachodniopomorski Uniwersytet Technologiczny
71-210 Szczecin ul. Żołnierska 49

Modelowanie 3D twarzy w systemach bezpieczeństwa publicznego

Słowa kluczowe: biometria, rozpoznawanie twarzy, systemy zarządzania tożsamością, bezpieczeństwo publiczne.

1. Wprowadzenie

Stale narastające w obecnych czasach zagrożenie atakami terrorystycznymi na świecie, duża mobilność i międzynarodowy zakres działania terrorystów i grup przestępczych implikuje potrzebę podjęcia odpowiednich działań w celu zwiększenia poczucia bezpieczeństwa obywateli. Działania prewencyjne policji i wszelkich służb specjalnych polegające na wzmożonej kontroli dworców, terminali lotniczych, przejść granicznych, dróg a nawet miejsc i obiektów, gdzie odbywają się imprezy masowe są bardzo uciążliwe dla obywateli. Zazwyczaj powodują skutek odwrotny od zamierzonego i wywołują głosy krytyki i oburzenia ze strony różnego rodzaju „obrońców wolności i praw człowieka". Idealnym w takich wypadkach rozwiązaniem byłoby wykorzystanie technicznych środków wsparcia w postaci bezinwazyjnych automatycznych systemów nadzoru i identyfikacji. Gwałtowny rozwój technologii i skuteczności zastosowań biometrycznych metod kontroli tożsamości potwierdzony w praktyce chociażby przez stosowany obecnie na całym świecie Zintegrowany System Automatycznej Identyfikacji Daktyloskopijnej IAFIS [1] stał się istotnym impulsem do prowadzenia w szerokim zakresie badań nad ulepszaniem i rozwojem nowoczesnych metod zarządzania tożsamością, badań, wykorzystujących najnowsze osiągnięcia technologii biometrycznych. Jedną z najmniej inwazyjnych, a jednocześnie skutecznych i najszybciej rozwijających się obecnie technik jest rozpoznawanie i identyfikacja człowieka oraz emocji na podstawie zdjęcia jego twarzy. Technika ta wymaga jednak ciągłej obserwacji twarzy monitorowanej osoby [2]. Rozpoznawanie twarzy znajduje zastosowanie nie tylko w aplikacjach sektora bezpieczeństwa, ale także może być pomocne np. w informatycznych systemach indywidualnego nauczania. O ile przy wykorzystaniu pojedynczego stanowiska komputerowego identyfikacja studenta może odbywać się na podstawie loginu i hasła, to już w systemach online może być bezwzględnie konieczna i stanowić element systemu rozpoznawania reakcji behawioralnych [3]. Możliwość

rozpoznawania poszczególnych osób w grupie oraz ich reakcji dzięki wykorzystaniu modelowana twarzy 3D może umożliwić w przyszłości również rozbudowę e-learningu dla większych grup studentów, jako część systemu zbiorowego nauczania w dużych systemach.

O pokładanych w tę technikę nadziejach świadczyć może ogrom nakładów finansowych i udział w wielu programach badawczych i użytkowych na całym świecie. Można tu wspomnieć np. o NGI [4] (Next Generation Identification) stanowiącym system Identyfikacji Nowej Generacji Departamentu Sprawiedliwości Stanów Zjednoczonych prowadzony przez FBI Federalne Biura Śledcze i realizowany przez potentata sektora technologii militarnych firmę SAFRAN przy wsparciu naukowym Uniwersytetu Zachodniej Wirginii (West Virginia University).

Również w Europie jest to temat bardzo istotny w obecnym czasie świadczy o tym zakres programów finansowanych przez UE, takich jak 7PR (7 Program Ramowy). Wymieniono w nim kilkanaście przedsięwzięć badawczo wdrożeniowych z zakresu bezpieczeństwa obywateli. Jednym z ciekawszych, wartych wymienienia, jest program INDECT [5] realizowany przez Akademię Górniczo-Hutniczą w Krakowie przy współudziale Policji Polskiej i Irlandzkiej oraz kilkunastu innych uczelni z terenu Polski i Europy. W innym programie UE Innowacyjna Gospodarka finansowanym z funduszy Europejskiego Programu Rozwoju Regionalnego realizowany jest SmartMonitor [6] przy wsparciu naukowym Zachodniopomorskiego Uniwersytetu Technologicznego w Szczecinie [7].

Głównym problemem w praktycznej realizacji identyfikacji osób na podstawie twarzy, oprócz zmiennych warunków oświetlenia i innych zakłóceń wywołanych ruchem osoby, jest brak możliwości uzyskania stałego obrazu twarzy [8]. W rzeczywistych warunkach, zwłaszcza przy obserwacji dynamicznej, nie ma możliwości rejestracji twarzy en face. O ile temat wykrywania i śledzenia ruchu obiektów na scenie w nowoczesnych systemach monitoringu wizyjnego jest przedmiotem wielu interesujących publikacji i realizacji praktycznych, to sama twarz nie jest jeszcze problemem dobrze poznanym [9]. Identyfikacja tożsamości osoby obserwowanej pod różnymi kątami często przekraczającymi dopuszczalne dla większości aplikacji normy, niejednokrotnie, przy częściowym zasłonięciu jej istotnych elementów, jest sytuacją najczęściej występującą i sprawiającą najwięcej problemów.

W związku z tym skutecznym rozwiązaniem mogłaby być identyfikacja osoby na podstawie jej modelu obróconego do pozycji identycznej jak rozpoznawana twarz.

Z uwagi na ten fakt, postanowiono wygenerować modele trójwymiarowe twarzy, a następnie sprawdzić możliwości rozpoznawania i identyfikacji osób na podstawie takiego modelu oraz porównać z rzeczywistą fotografią twarzy.

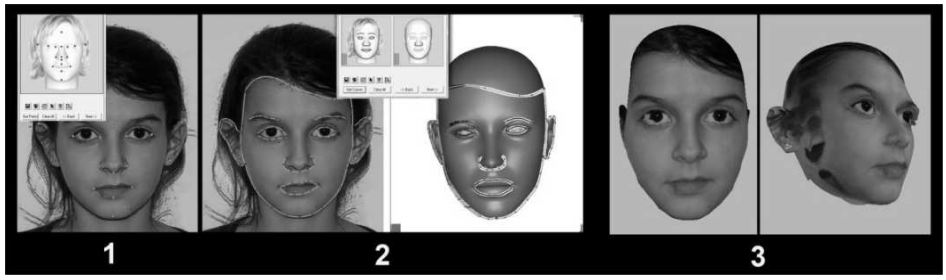
2. Modelowanie 3D twarzy

Z uwagi na wysokie wymagania sprzętowe ograniczeniem w praktycznych zastosowaniach aplikacji do generowania modelu 3D jest najczęściej niska jakość i mała dokładność modelu, daleko odbiegająca od oryginału, jak również „płaski" model pozbawiony szczegółów, zwłaszcza w programach korzystających tylko z jednego zdjęcia en face.

Mimo słabych efektów końcowych czas renderowania fotografii jest i tak dość długi, nawet powyżej 20 minut. W celu wyboru optymalnej metody zbadano możliwości kilku aplikacji a najciekawsze wyniki przedstawiono poniżej. Przy wyborze aplikacji do testów kierowano się przede wszystkim możliwościami i jakością otrzymywanych modeli, jednak nie bez znaczenia był również jej koszt.

FaceShop 3.15

Jednym z testowanych programów była aplikacja FaceShop 3.15 w 15 dniowej wersji Trial. Po wprowadzeniu wybranego obrazu, wskazaniu punktów charakterystycznych rys. 1 pkt.1 oraz obwiedni głównych elementów twarzy rys. 1 pkt. 2 program generuje model trójwymiarowy twarzy, który można dowolnie obracać.



Rys. 1. Etapy lokalizacji cech pkt. 1 i 2 oraz wynik końcowy renderowania pkt. 3 w programie FaceShopPro 3.15 (źródło – opracowanie autorskie)

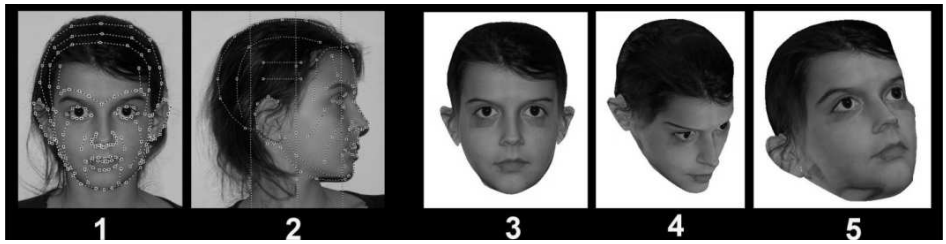
Jak widać na załączonym rysunku nr 1 pkt. 3, model en face jest w miarę dokładny, jeśli jednak przyjrzymy się twarzy pod różnymi kątami widać wiele bardzo istotnych odstępstw od oryginału, co czyni go nieprzydatnym do dalszych badań.

LOOXIS Faceworx

Innym programem do modelowania 3D twarzy jest aplikacja LOOXIS Faceworx niemieckiej firmy LOOXIS GmbH. Po wczytaniu 2 fotografii (en face i profil lewy) należy ustawić linie obwiedni głównych elementów twarzy i linie pomocnicze. Program wymaga dużego nakładu pracy przy wytyczaniu wskaźników, dodatkowo linie można dowolnie modyfikować i zmieniać ilość punktów krzywizny. Jest to dość żmudna praca i wymaga wręcz artystycznych zdolności, gdyż prawidłowego przebiegu część linii

pomocniczych można się tylko domyślać. W efekcie otrzymujemy model o podobnych właściwościach jak w poprzednio prezentowanym programie FacShopPro rysunek 2.

Jest on niedokładny i posiada wiele błędów, a profil twarzy jest za każdym razem niepowtarzalny i zależy wyłącznie od zdolności rysunkowych operatora, co czyni go nieprzydatnym do dalszych badań.



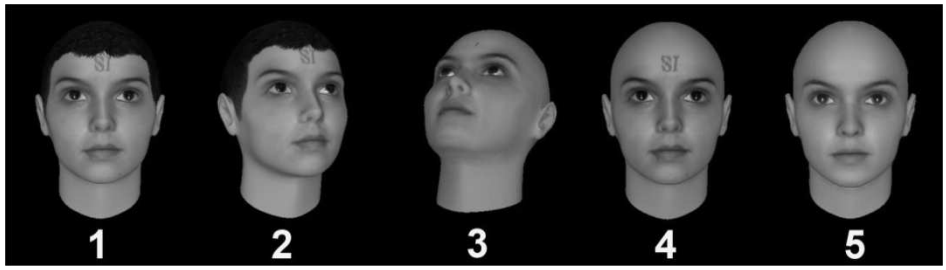
Rys. 2. Etapy lokalizacji linii pomocniczych pkt. 1 i 2 oraz wynik końcowy renderowania pkt. 3, 4, 5 w programie Looxis Faceworks (źródło – opracowanie autorskie)

FaceGen Modeller

Jedną z ciekawszych dostępnych na rynku aplikacji, umożliwiających generowanie modeli 3D głowy, na podstawie odpowiednich zdjęć twarzy, jest program FaceGen Modeller. Oprogramowanie w wersji FaceGen Modeller 3.5 Fre jest dostępne bezpłatnie. Jedynym minusem wersji darmowej jest znak SI pozostawiany na czole modelu. Jednak dla celów badania okazał się on nieistotny, gdyż przy rezygnacji z tekstury obiektu znika on prawie całkowicie. Ponadto do dalszych testów przewidziano wykorzystanie części twarzy zawierające charakterystyczne cechy antropologiczne od linii brwi do podbródka. Z uwagi na powyższe wykonania modeli wybrano aplikację FeceGen Modeller. Posiada ona ponadto bardzo rozbudowane możliwości generowania i modyfikacji modelu 3D, co okazało się istotne w dalszej części pracy. Tworzenie modelu odbywa się na podstawie 3 fotografii głowy on face, z prawego i lewego profilu. Na każdej z fotografii konieczne jest ręczne wybranie i zaznaczenie podstawowych charakterystycznych punktów biometrycznych wyznaczających między innymi szerokość i wysokość nosa, ust, twarzy, centra oczu, itd. Ponieważ są to dość charakterystyczne punkty, a obrabiany obraz można powiększać wskazanie właściwego umiejscowienia punktów nie stanowi większego problemu rysunek 3.

Rys. 3. Zdjęcia wykorzystane do tworzenia modelu 3D, z wyznaczonymi punktami charakterystycznymi (źródło – opracowanie autorskie)

Do wyznaczonych punktów charakterystycznych dopasowywany jest trójwymiarowy model i wprowadzone zdjęcia twarzy. Następnie na model „naciągane" są 3 zdjęcia en face i profili z uwzględnieniem lokalizacji elementów twarzy wskazanych w postaci wyznaczonych punktów przez użytkownika. Tak przygotowany model twarzy można obracać pod dowolnym kątem w osi pionowej i poziomej, korygować większość parametrów geometrii i wyrazu twarzy. Wykonawcy aplikacji poszli jeszcze dalej i przygotowali możliwość zmiany według gotowych schematów: rasy, płci, wieku, emocji, tekstury powierzchni skóry, itp.



Rys. 4. Zdjęcia nr 1, 2, 3, 4 przedstawiają rzuty pod różnymi kątami otrzymanego modelu 3D głowy z naniesiona teksturą fotografii twarzy. Zdjęcie nr 5 przedstawia ten sam model bez tekstury. (źródło – opracowanie autorskie)

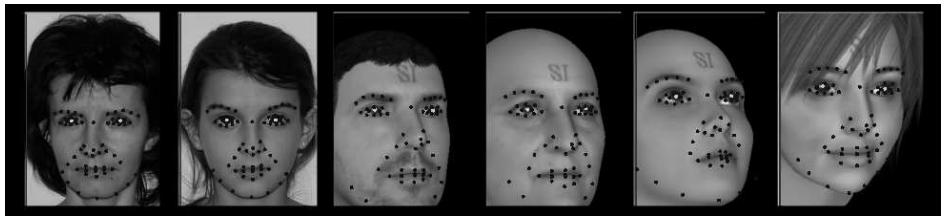
Jak widać na rysunku nr 4 – otrzymany wynikowy trójwymiarowy model jest bardzo podobny do osoby na zdjęciach.

3. Weryfikacja modelu 3D i twarzy wzorcowej

W celu weryfikacji subiektywnych odczuć podobieństwa otrzymanych modeli do twarzy wzorcowej postanowiono porównać lokalizacje punktów charakterystycznych twarzy. Do dokładnego wyznaczenia i analizy cech biometrycznych sprawdzono działanie różnych programów umożliwiających lokalizacje punktów charakterystycznych.

Luxand FaceSDK 1.7.

Komercyjna aplikacja FaceSDK 1.7 firmy Luxand umożliwia według zapewnień producenta detekcje wielu twarzy na zdjęciu lub z kamery pod kątem od -30 do 30 stopni w osi obrotu oraz od -30 do 30 stopni poza osią obrotu, z szybkością od 0,05 s do 1,1s. Na twarzy wskazuje do 40 punktów charakterystycznych zawierających między innymi: oczy, brwi, usta, nos, twarz, kontur pod katem od -30 do 30 stopni w osi obrotu oraz od -10 do 10 stopni poza osią obrotu, z szybkością 0,9 s.



Rys. 5. Punkty charakterystyczne twarzy wskazane w programie Luxand FaceSDK 1.7 (źródło – opracowanie autorskie)

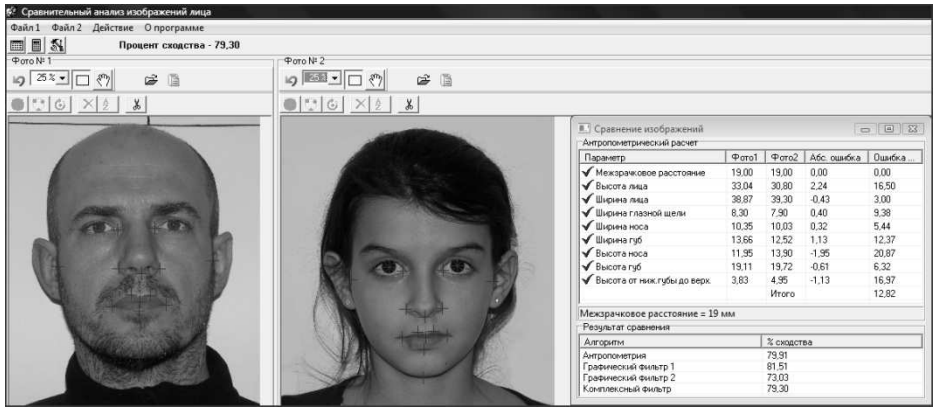
W darmowej wersji demonstracyjnej nie ma możliwości porównania portretów, odczytu współrzędnych ani korekcji położenia błędnie zlokalizowanych punktów. Jak widać na rysunku nr 5 (brak możliwości powiększenia uzyskanego obrazu) program działa dobrze przy twarzach na wprost przy odchyleniach od osi pojawiają się błędy. Z uwagi na powyższe aplikacja nie została wykorzystana do dalszych badań.

Client Portret 5.0

Aplikacja Client Portret 5.0 firmy PortLand Ltd. w darmowej wersji demonstracyjnej umożliwia porównanie dwóch portretów twarzy na wprost. Po wprowadzeniu obrazu program wykrywa automatycznie 18 punktów charakterystycznych twarzy, położenie błędnie wykrytych punktów można korygować ręcznie.

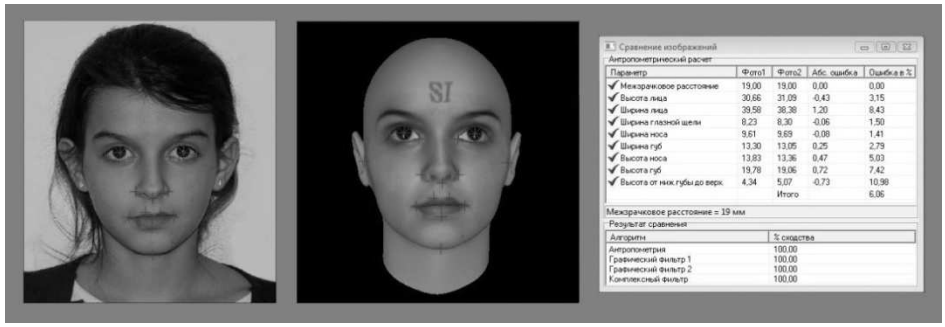
Punkty służą do wyliczenia rozmiarów 9 elementów głównych twarzy tj.:

- rozstawu oczu,
- wysokości twarzy,
- szerokości twarzy,
- szerokości otworu oczu,
- szerokości nosa,
- szerokości ust,
- wysokości ust.



Rys. 6. Porównanie dwóch portretów twarzy w aplikacji Client Portret 5.0 (źródło – opracowanie autorskie)

Portrety porównywane są na podstawie różnicy wymiarów 9 podstawowych elementów twarzy natomiast sam obraz przy użyciu 2 filtrów graficznych. Wyniki podawane są, jako procentowa rozbieżność między uzyskanymi w ten sposób danymi antropometrycznymi. Wynik końcowy podawany jest, jako procentowy stopień podobieństwa rysunek 5 (okno prawe, w tym wypadku 79,3 %). W zależności od równomierności oświetlenia rozdzielczości i jakości fotografii program z większą lub mniejszą dokładnością wyszykuje automatycznie, na obu porównywanych zdjęciach charakterystyczne punkty biometryczna, mierzy odległości i oblicza procentowy stopień podobieństwa. W wypadku błędnych wskazań rozmieszczenie punktów można korygować ręcznie. Rezultaty porównania modelu 3D ze zdjęciem wzorcowym twarzy przedstawiono na rys. 7.



Rys. 7. Lokalizacja podstawowych punktów antropometrycznych modelu z twarzą na zdjęciu oraz porównanie wymiarów w aplikacji Client Portret 5.0 (źródło – opracowanie autorskie)

Wyniki pomiarów wskazują na pewne rozbieżności między pierwowzorem a modelem, jednak w końcowym podsumowaniu zawierają się w tolerowanym zakresie i potwierdzają bardzo wysokie podobieństwo obu fotografii:

- antropometria – 100%,
- filtr graficzny nr 1 – 100%,
- filtr graficzny nr 2 – 100%,
- filtr kompleksowy – 100%.

Otrzymane wyniki dają podstawę do tworzenia nowych modeli i kontynuowania dalszych badań w tym zakresie.

4. Rodzinna baza twarzy (Base family)

Jak już wcześniej wspomniano do stworzenia trójwymiarowego modelu głowy potrzebne są 3 zdjęcia twarzy. Z uwagi na brak dostępu do bazy odpowiednio wykonanych fotografii rodziców z dziećmi, do testów oprogramowania przygotowano bazę autorską zdjęć przypadkowo dobranych rodzin. Do przeprowadzenia koniecznych testów wykonano dla każdej osoby po 3 zdjęcia twarzy w rzutach en face, lewy oraz prawy profil. Przy użyciu programu FaceGen Modeller 3.5 Free wytypowanego w poprzednim etapie, wygenerowano modele trójwymiarowe twarzy. Wszystkie modele pogrupowano względem rodzin, przyjmując 4-ro cyfrowe oznaczenia liczbowe, gdzie, np. 0011 oznacza rodzina nr 001, płeć męska, ojciec.

W szczegółowym rozbiciu 0011: 001 – 3 pierwsze cyfry są nr rodziny, ostatnia cyfra pełni wieloraką rolę, 1 i 2 – rodzice a od 3 do 9 - dzieci, przy czym oznaczeniem płci męskiej są cyfry nieparzyste i a contrario cyfry parzyste oznaczają płeć żeńską, oznacza ona również kolejność narodzin. Otrzymaną bazę 3D modeli twarzy przedstawiono na rysunku 8.

Rys. 8. Zdjęcia modeli 3D twarzy pozbawione tekstury w bazie family (źródło – opracowanie autorskie)

5. Analiza otrzymanych modeli

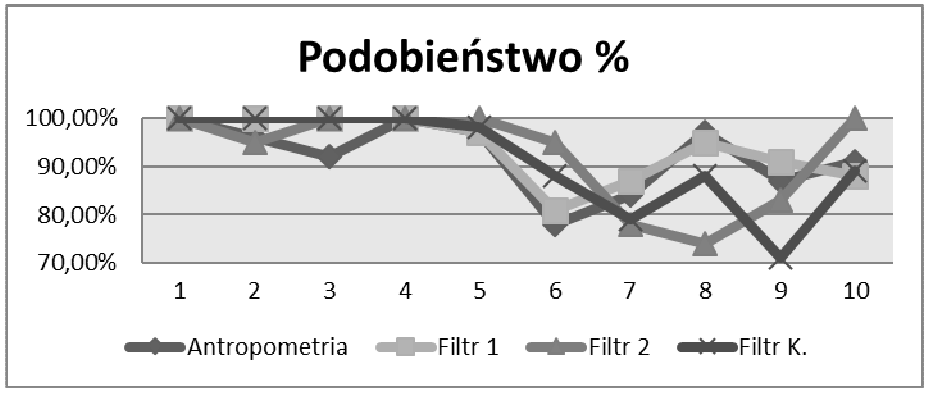
Analizując po rożnymi kątami fotografie otrzymanych w wyniku obróbki modeli 3D na pierwszy rzut oka widać bardzo duże podobieństwo z wzorcowymi zdjęciami twarzy. Wyniki porównania danych antropometrycznych w aplikacji Client Portret 5.0. potwierdzają odczucia subiektywne tabela nr 1. Przeanalizowano 10 kolejnych modeli bazy family – Rysunek 9. Punkty charakterystyczne wskazane zostały ręcznie, gdyż

aplikacja nie radziła sobie zupełnie z ich lokalizacją na modelu twarzy. W niektórych przypadkach widać drobne odstępstwa od modelu. Istotny wpływ na wyniki ma tu zapewne duża rozpiętość wiekowa osób umieszczonych w bazie. Z uwagi na sposób formowania modelu na podstawie zdjęć, a więc wyglądu twarzy zwłaszcza skóry zauważono silny wpływ na efekt końcowy zewnętrznych oznak starzenia się, zmarszczek, blizn, nadmiernej otyłości, czy szczupłości.

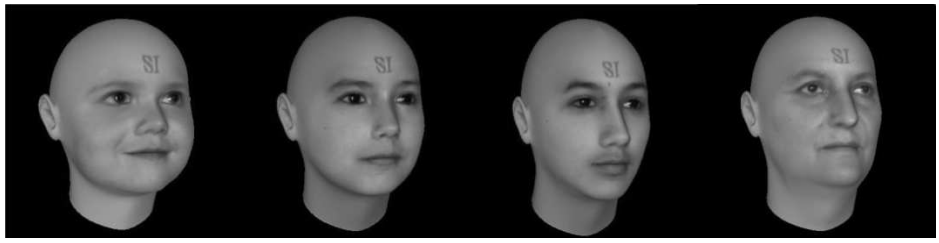
Tabela 1. Porównanie danych antropometrycznych zdjęcia bazowego (fot. 1) i otrzymanego modelu twarzy (fot. 2) dla pozycji 0011 na podstawie w wyniku otrzymanego w aplikacji Client Portret 5.0

OBLICZENIA ANTROPOMETRYCZNE				
Parametr	**fot. 1**	**fot. 2**	**Błąd abs.**	**Błąd %**
Rozstaw źrenic	19,00	19,00	0,00	0,00
Wysokość twarzy	30,66	31,09	-0,43	-1,40
Szerokość twarzy	39,58	38,38	1,20	3,03
Szerokość oka	8,23	8,30	-0,07	-0,85
Szerokość nosa	9,61	9,69	-0,08	-0,83
Szerokość warg	13,30	13,05	0,25	1,88
Wysokość nosa	13,83	13,36	0,47	3,40
Wysokość warg	19,78	19,06	0,72	3,64
Wysokość ust	4,34	5,07	-0,73	-16,82
	Razem			7,96

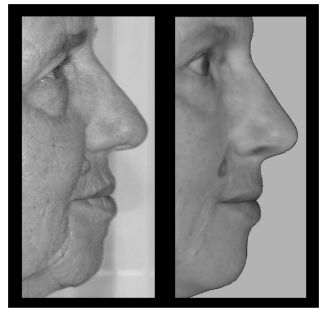
WYNIK PORÓWNANIA	**PODOBIEŃSTWO**
Antropometria	100 %
Filtr graficzny 1	100 %
Filtr graficzny 2	100 %
Filtr kompleksowy	100 %



Rys. 9. Wyniki podstawowych parametrów dla 10 modeli z bazy family

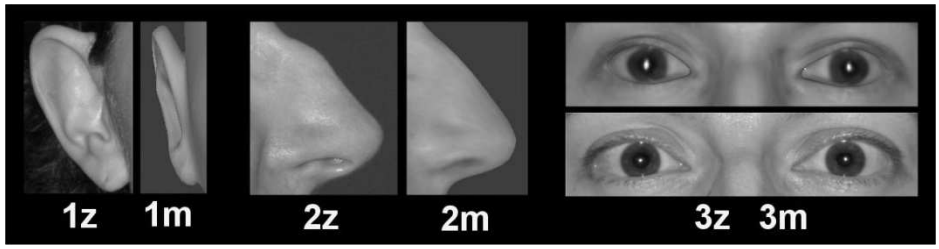


Rys. 10. Zdjęcie przedstawia modele twarzy osób w wieku 2, 8, 12 i 55 lat



Rys. 11. Od lewej zdjęcie osoby w wieku 65 lat i jej model z prawej

Wyjątkowo intensywne zmiany będące wynikiem procesu formowania kształtu twarzy dziecka w pierwszych latach życia, ale widoczne także w okresie pokwitania, aż do późnego nastolatka powodują duże problemy dla większości programów opartych na modelu twarzy człowieka dorosłego. Optymalnym obiektem badań, generującym w tym wypadku najmniejsze błędy, jest twarz człowieka, w wieku 20-50 w pewnych warunkach osobniczych nawet 15 - 60 lat (rysunek 10). Zmiany zachodzące w wyglądzie zewnętrznym w późnym okresie życia, zwłaszcza u osób w podeszłym wieku 75-80, są już na tyle zaawansowane, że uniemożliwiają w prosty sposób automatyczne odtworzenie wcześniejszego wyglądu (rysunek 11). Innym zauważalnym problemem jest zbyt mała liczba punktów charakterystycznych znakowanych na fotografii powodująca ograniczenia w dokładności doboru trójwymiarowego model twarzy, wykorzystywanego następnie, jako szablon do nałożenia zdjęcia lub tekstury z bazy programu. Uwidacznia się to, jako tendencja automatycznego uśredniania wyglądu, zaokrąglania i wygładzania ostrych konturów twarzy osób o nietypowej urodzie.



Rys. 12. Błędy modelowania. Z lewej zdjęcie, z prawej fotografia odpowiadającego mu modelu ucha pkt. 1, nosa pkt. 2 oraz oczu wraz z oprawą pkt. 3 (na dole zdjęcie, u góry otrzymany model) (źródło – opracowanie autorskie)

6. Podsumowanie

Otrzymane wyniki zachęcają do prowadzenia dalszych badań w zakresie modelowania 3D twarzy. Jak widać z przeprowadzonych testów, jest to jedna z metod rozwiązania problemu nieinwazyjnej identyfikacji osób na podstawie twarzy w systemach ochrony bezpieczeństwa publicznego.

Bibliografia

1. IAFIS (ang. Integrated Automated Fingerprint Identification System) Zintegrowany System Automatycznej Identyfikacji Daktyloskopijnej Federalnego Biura Śledczego FBI (ang. Federal Bureau of Investigation), http://www.fbi.gov/about-us/cjis/fingerprints_biometrics/iafis/iafis, dostęp 20.04.2013r.
2. Wosiak S., Buda K., Wiliński A., Poszukiwanie twarzy i rozpoznawanie emocji w scenach zmiennych, Metody Informatyki Stosowanej, 5/2011, str. 147-159.
3. Wosiak S., Wzorce behawioralne w środowisku akademickim szansą na personalizację procesów edukacji, Polskie Stowarzyszenie Zarządzania Wiedzą, Studia i Materiały, nr 57, 2011, str. 298-311.
4. System Identyfikacji Nowej Generacji NGI (Next Generation Identification), http://www.fbi.gov/about-us/cjis/fingerprints_biometrics/ngi, dostęp 14.04.2013r.
5. INDECT (ang. Intelligent information system supporting observation, searching and detection for security of citizens in urban environment) - Inteligentny system informacyjny wspierający obserwację, wyszukiwanie i detekcję dla celów bezpieczeństwa obywateli w środowisku miejskim. Program ramowy UE obszar obronności i bezpieczeństwa, 7 PR, nr kontraktu: 218086, http://www.indect-project.eu/, dostęp 06.04.2013r.

6. SmartMonitor, Budowa prototypu innowacyjnego systemu bezpieczeństwa opartego o analizę obrazu, Program Operacyjny Innowacyjna Gospodarka 2007-2013 finansowany z Europejskiego Funduszu Rozwoju Regionalnego (EFRR) - stan na 3 lipca 2011 r. na podstawie KSI SIMIK 07-13, Projekt: UDA-POIG.01.04.00-32-008/10-01.
7. Frejlichowski D., Gościewska K., Forczmański P., Nowosielski A., Hofman R., SmartMonitor: recent progress in the development of an innovative visual surveillance system, Journal of Theoretical and Applied Computer Science Vol. 6, No. 3, 2012, str. 28–35, http://www.jtacs.org
8. Forczmanski P., Kukharev G. A., Shchegoleva N. L., «An algorithm of face recognition under difficult lighting conditions» - Electrical Review, 2012, nr 10B, str. 201-204.
9. Forczmański P., Frejlichowski D., Nowosielski A., Hofman R., Aktualne trendy w tworzeniu systemów inteligentnego monitoringu wizyjnego, Metody Informatyki Stosowanej Nr 4/2011 (29), str. s. 19-32, ISSN 1898-5297.

Streszczenie

Niniejszy artykuł porusza temat aktualnych tendencji rozwojowych i praktycznego wykorzystania osiągnięć biometrycznych technologii identyfikacji osób w systemach informatycznego wsparcia bezpieczeństwa publicznego. Przedstawiono jeden z głównych problemów, jakim jest identyfikacja twarzy osób w warunkach rzeczywistych oraz podjęto próbę praktycznego rozwiązania tego problemu w postaci modelowania 3D twarzy. W przyjętej koncepcji nadzór prewencyjny ze strony służb za to odpowiedzialnych zostałby częściowo zautomatyzowany i wsparty systemem informatycznym identyfikującym osoby na podstawie wzorcowych modeli 3D uzyskanych drogą renderowania trójwymiarowego zdjęć wzorcowych twarzy. Zakłada się, że zdjęcia twarzy pozyskiwane byłyby za pomocą systemu monitoringu wizyjnego.

Abstract

The paper concerns the subject of current developments and achievements of the practical use of biometric identification technology in computer systems for public safety support. It presents a major problem, which is to identify faces of people in “a real” and practical attempts to solve this problem in the form of 3D face modeling. The approach may be used by the preventive supervision departments responsible and may be partially automated. An information may be supported by the system that identifies people based on standard 3D models obtained by rendering three-dimensional images of best face. It is assumed that the face images would be obtained through video monitoring system.

Keywords: biometrics, face recognition, identity management systems, public safety

Przemyslaw Makiewicz
Krzysztof Penkala
Department of Systems, Signals and Electronics Engineering,
Faculty of Electrical Engineering,
West Pomeranian University of Technology,
Sikorskiego 37
70-313 Szczecin
Poland

Virtual magnetic resonance imaging as a didactic aid

Keywords: NMR, MRI, VMRI, spin echo method, SE, computer simulation, didactics

1. Introduction

Magnetic Resonance Imaging

Images obtained by Nuclear Magnetic Resonance (NMR) are the graphical representation of the distribution of certain physical properties of the object. In practice, Magnetic Resonance Imaging (MRI) is based on the nuclei of hydrogen atoms. They have half-spin, so they are susceptible to NMR phenomena. Furthermore, there are commonly present in the human body [1]. This allows identifying three parameters characterizing studied tissues that determine MRI contrast obtained [2]:

- PD - the density of protons (proton density, ρ). This is the amount of nuclei (due to the use of hydrogen nuclei, the term density of protons has been used) that are subject to the phenomenon of resonance, in a unit volume;
- T1 - time of longitudinal relaxation. This is the time constant describing the rate of re-growth of the longitudinal magnetization component of the atomic nucleus;
- T2 - transverse relaxation time. This is the time constant describing how rapidly the component of transverse magnetization disappears.

The signal carrying information about the above mentioned properties of the test subject is obtained in response to the specific excitation sequence. It consists of radio frequency (RF) pulses that change the magnetization vector of the particles susceptible to the NMR phenomenon. Reading of the signal (echo) from the currently projected section is made possible by controlling the magnetic field gradient and excitation sequence

parameters. The use of gradients allows choosing the appropriate cross-section [3,4]. By manipulating the excitation sequence, contrast can be weighted by various tissue parameters (PD, T1 and T2). The spin echo sequence is presented on Figure 1. It can be described by two parameters:

- TR - repetition time. This is the time between the RF pulse bursts;

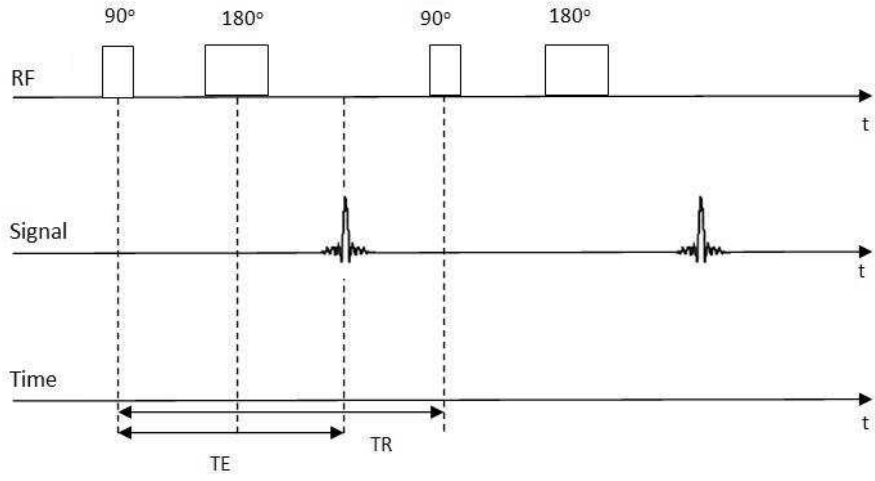


Figure 1. RF pulse sequence for the SE method

- TE - echo time. This is the time after which the signal is read (echo).

The use of SE sequence allows obtaining an image in which the contrast is weighted by any of the parameters of the tissue. Depending on the values of TR and TE, the image shows the distribution of PD, T1 or T2. The impact of individual parameters on the imaging results describes a mathematical formula available in the literature [5]:

$$S = \rho(1 - e^{-\frac{t}{T_1}})e^{-\frac{t}{T_2}} \tag{1}$$

Virtual MRI

Virtual Magnetic Resonance Imaging is carried out by means of computer data processing which determines the contrast obtained. This way, it is possible to obtain an image of any desired contrast. VMRI achieves this effect by performing appropriate calculations, without the use of an MRI scanner and involvement of the patient.

The idea of developed VMRI technique is presented on Figure 2 [6]. The first block contains the data necessary to create the model of the patient. These should include data for three images (PD, T1 and T2 weighted contrasts) as well as TR and TE parameters for each of three pictures. This information is saved as DICOM data, and reading them

does not constitute any problem. Mathcad software was used for VMRI simulation in our research.

When processing previously obtained images with the VMRI technique, other problem occurs. Due to the high cost of the MRI examinations, generally images of the same cross-section are not recorded in all three contrasts. Skipping images with the lowest diagnostic value for a given medical purpose can reduce time and thus keep the cost at appropriate level. Unfortunately, it prevents full use of the VMRI technique. However, analysis of some cases in our previous studies (neurology, orthopedics) allowed determining that it is possible to omit the missing image at the stage of creating a model of the patient. This topic requires further research on optimization of the VMRI technology [6,7].

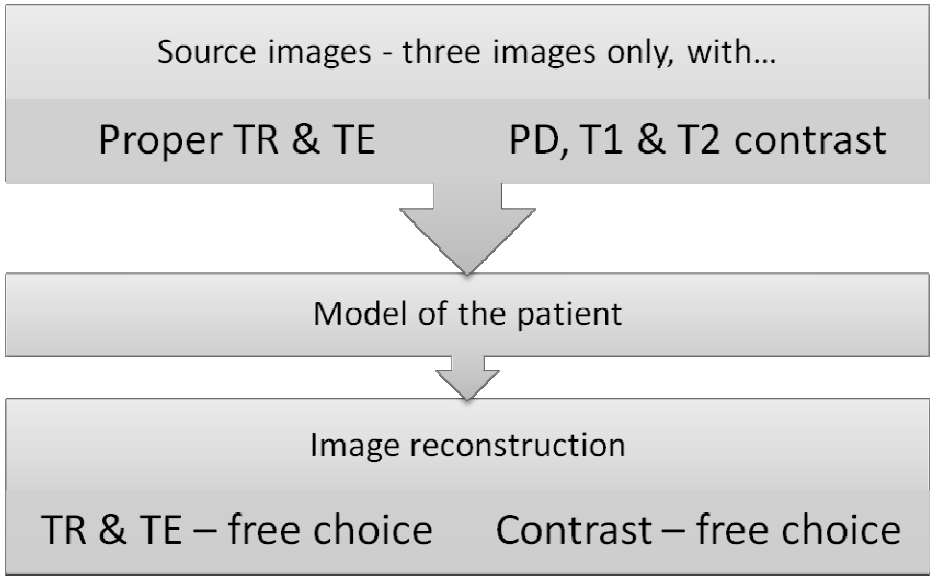


Figure 2. Idea of VMRI

The advantage of VMRI simulation is the ability to generate images in real-time and without involving the equipment and the patient. The user is not limited in any way in selection of the excitation sequence timing parameters. It is an exciting prospect for both the student who wants to better understand the rules governing the MRI, and the doctor who wants to examine the patient as accurately as possible.

Use of the VMRI simulation for teaching purposes allows continuous changing of the TR and TE parameters, in order to observe their effect on obtained contrast. For example, a student can observe the result of simulation for a short TR time and long TE time. Such sequence is never applied in clinical practice. This is justified by the fact that with such set of parameters, it is not possible to obtain a good contrast. In the picture, a

significant effect of all three parameters of the tissue will be present. As a result, the image is of little diagnostic value. With the simulation, students can easily see that in fact such set of parameters is useless. Figure 3 [8,9] shows image generated with VMRI technique. Such images have both didactical and diagnostic value.

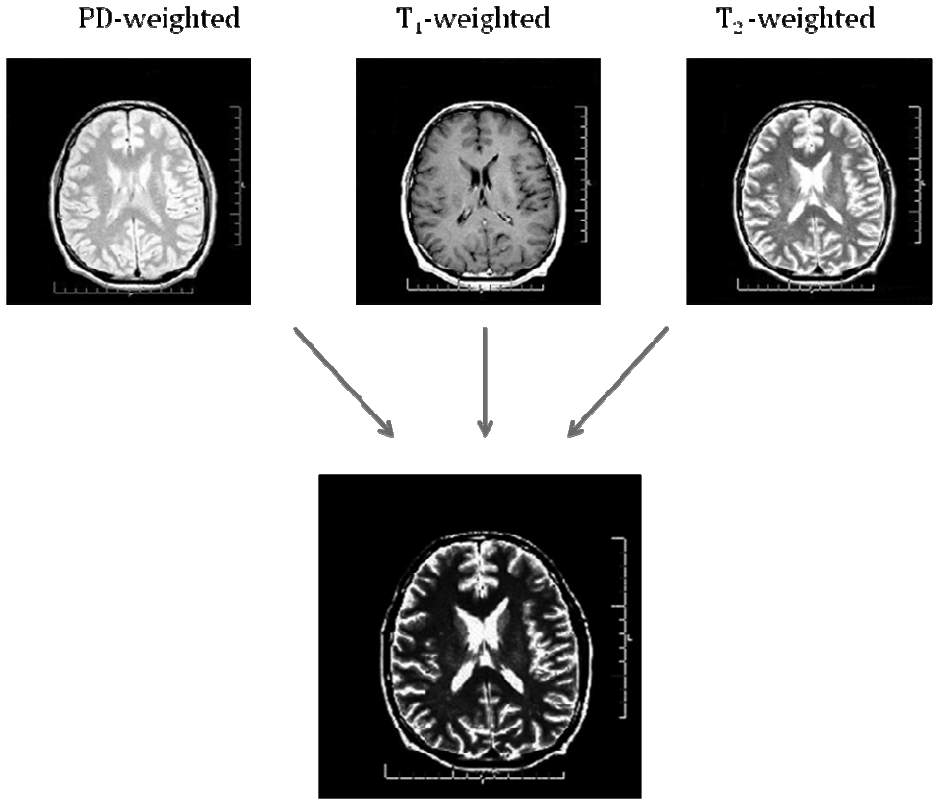


Figure 3. Example of VMRI generated image (TR = 25 s, TE = 1,25 s)

2. Conclusions

The aim of this study was to develop a method for creating a model or a digital phantom of the patient on the basis of three MRI images. After reading the source materials, the VMRI system automatically calculates the values that produce a phantom of the patient. Then the simulation tool can generate images from the model, for any selected time parameters of the excitation sequence.

The main expectation to the simulation was to demonstrate the impact of TR and TE parameters on MRI results. Simulation has the didactic value for those involved in the field of Biomedical Engineering. It allows much better way to introduce Magnetic

Resonance Imaging measures to the students studying technical fields. However, it is particularly important also for medical students. They are not familiar with analyzing equations as well as complicated diagrams, and draw conclusions from them. Therefore, the ability to manipulate imaging parameters and observe immediately, on-line, their effects on image on the computer screen, is a very useful teaching aid.

The value of VMRI simulation in teaching has been verified by the first author of the work. Within the framework of the Erasmus program, he spent a semester scholarship at the Fachhochsule Stralsund, Germany, where he presented simulations described in this study to the students of Biomedical Engineering. In consultation with the lecturer responsible for the subject of "Medical Imaging Systems", author of a series of lectures supplemented the activities carried out based on the simulation. All persons participating in the classes confirmed that VMRI description of the SE method makes it much easier to learn the principles of MRI as well as image contrast manipulation.

Recently, the method of VMRI has been included also in the course of “Medical Informatics and Telemedicine” at the Faculty of Electrical Engineering of the West Pomeranian University of Technology, Szczecin. It is a facultative course for students of Electronics and Telecommunications, with specialization in Electronic Systems.

It is also planned to develop a teaching platform for medical students from the Pomeranian Medical University in Szczecin. First tests carried out with the staff from the Department of Radiology of the PMU have demonstrated very good teaching outcomes.

References

1. Hausser K.H., Kalbitzer H.R. (1993) NMR w biologii i medycynie Badania strukturalne, tomografia, spektroskopia in vivo (in Polish). Wydawnictwo Naukowe UAM, Poznań.
2. Pruszyński B. et al. (2000) Diagnostyka obrazowa: podstawy teoretyczne i metodyka badań (in Polish). Wydawnictwo Lekarskie PZWL, Warszawa.
3. Weishaupt D., Köchli V.D., Marincek B. (2001) Wie funktioniert MRI Eine Einführung In Physik Und Funktionsweise der Magnetresonanzbildgebung. Springer-Verlag, Berlin.
4. Vlaardingerbroek M.T., den Boer J.A. (1996) Magnetic Resonance Imaging. Springer-Verlag, Berlin.
5. Gonet B. (1997) Obrazowanie magnetyczno-rezonansowe. Zasady fizyczne i możliwości diagnostyczne (in Polish). Wydawnictwo Lekarskie PZWL, Warszawa.
6. Makiewicz P., Penkala K., Lubiński W. (2011) Virtual Magnetic Resonance Imaging (VMRI) – innowacyjna technologia wspomagająca diagnostykę schorzeń nerwu wzrokowego (in Polish). Streszczenia. III Sympozjon Sekcji Neurookulistyki i Elektrofizjologii Klinicznej PTO: 103-105, Międzyzdroje.

7. Makiewicz P., Penkala K., Lubiński W., Walecka A. (2012) Virtual Magnetic Resonance Imaging (VMRI) as a novel technology supporting diagnostics of the optic nerve diseases. 50th ISCEV Symposium proceedings: 58-59, Valencia.
8. Makiewicz P., Penkala K. (2010) Symulacja wpływu parametrów sekwencji wzbudzającej na wynik obrazowania metodą echa spinowego w rezonansie magnetycznym (in Polish). Inżynieria Biomedyczna – ActaBio-Optica et Informatica Medica 16(2): 76-79.
9. Tadeusiewicz R. (2010) Wykład 4 z kursu Techniki Obrazowania Medycznego (in Polish). http://moodle.cel.agh.edu.pl/msib/mod/ re-source/view.php?id=253.

Abstract

The paper concerns aspects of using Nuclear Magnetic Resonance in medicine. The phenomenon of NMR is basis for the Magnetic Resonance Imaging (MRI). Simulation of creating an image in spin echo (SE) method of MRI is used to obtain additional, arbitrary chosen images of the patient. Virtual MRI technique allows changing contrast of images in real time without use of any special equipment. Despite diagnostic benefits, VMRI is very helpful in didactics. Performing of such simulations greatly improves understanding of MRI principles and image contrast manipulation by the students.

Streszczenie

Artykuł dotyczy aspektów zastosowań magnetycznego rezonansu jądrowego w medycynie Symulacja tworzenia obrazu echa spinowego (SE) w metodzie MRI jest używana w celu uzyskania dodatkowych, dowolne wybranych obrazów pacjenta. Wirtualna technika MRI pozwala na zmianę kontrastu obrazów w czasie rzeczywistym, bez użycia specjalnego sprzętu. Pomimo świadczeń diagnostycznych, VMRI jest bardzo pomocna w dydaktyce. Przeprowadzenie takich symulacji znacznie poprawia zrozumienie zasad MRI i manipulacji kontrastu obrazu przez uczniów.

Słowa kluczowe: NMR, MRI, VMRI, metoda echa spinowego SE, symulacje komputerowe, dydaktyka

Konrad Jędrzejewski
Wydział Elektroniki i Informatyki
Politechnika Koszalińska
Śniadeckich 2
75-453 Koszalin

Porównanie metod filtracji obrazów z kamer monitorujących

Słowa kluczowe: monitoring, filtracja, pikselizacja, filtr medianowy statystyczny, filtr otwierający.
Keywords: monitoring, filtering, pixelization, statistical median filter, the filter opening.

1. Wstęp

Wraz z wprowadzeniem kamer monitorujących do systemów bezpieczeństwa w latach 60. XX wieku, w systemach tego typu nastąpił przełom. Użycie kamer wideo w latach 60. i 70. szybko wzrastało z powodu zwiększenia ich niezawodności, obniżenia kosztów i technologicznych ulepszeń w kamerach opartych na lampach analizujących. Najbardziej znaczącym postępem w technologii wideo w latach 80. było wynalezienie i wprowadzenie kamer półprzewodnikowych. Lata 90. ujrzały integrację technologii komputerowej z technologią kamer bezpieczeństwa [1]. W latach 80. usprawniono system bezpieczeństwa poprzez dodanie wykrywacza ruchu (VMD), który dokonywał analizy obrazu z kamery i informował o ruchu na obszarze monitorowania.

Do końca lat 80. wykrywacz ten był urządzeniem analogowym (AVMD). Analogowy wykrywacz ruchu nadawał się jedynie do stosowania w pomieszczeniach, gdzie poziom światła mógł być kontrolowany. W przypadku wykorzystania AVMD na zewnątrz, często występowały fałszywe alarmy czyniące system bezpieczeństwa bezużytecznym. W latach 90. wprowadzono cyfrowy wykrywacz ruchu (DVMD). Urządzenie to okazało się skuteczne do stosowania w pomieszczeniach i na zewnątrz. Dziś ciężko wyobrazić sobie skuteczny system bezpieczeństwa bez użycia kamery.

Systemy bezpieczeństwa służą do monitorowania fabryk, przedsiębiorstw, sklepów. Są szeroko stosowane nie tylko do wykrywania ruchu, ale także do ostrzegania o podejrzanych zachowaniach, torbach pozostawionych na dworcach, ruchu w nieodpowiednim kierunku (na przykład na autostradach). Potrafią nie tylko wykryć zagrożenie, ale także poddać je analizie i w razie potrzeby odpowiednio zareagować. Współczesne kamery przemysłowe zaopatrzone są najczęściej w wykrywacz ruchu

umieszczony w obudowie razem z kamerą. W zależności od specyfiki miejsca, w którym mają zostać wykorzystane, stosuje się inne typy kamer. Przemysłowe kamery zaopatrzone są często w reflektor składający się z diod podczerwonych, aby mogły monitorować także w trybie nocnym. Kamery mogą działać w sieci, gdzie każda posiada swój adres IP. Możliwe jest wydzielenie obszaru, w którym system ma wykrywać ruch. Jeżeli obraz z kamery obejmuje na przykład kawałek ruchliwej ulicy, można wyłączyć ją z obszaru monitorowania (obraz nadal będzie rejestrowany w całości). Ważną cechą systemów bezpieczeństwa jest możliwość kalibracji, czyli określenia czułości systemu na ruch. Obraz w systemach monitoringu nagrywany jest często dopiero w momencie wykrycia ruchu. Stosowanie wykrywacza ruchu w każdej kamerze osobno pozwala uniknąć gromadzenia się ogromnej ilości danych, zwłaszcza jeśli kamer jest dużo.

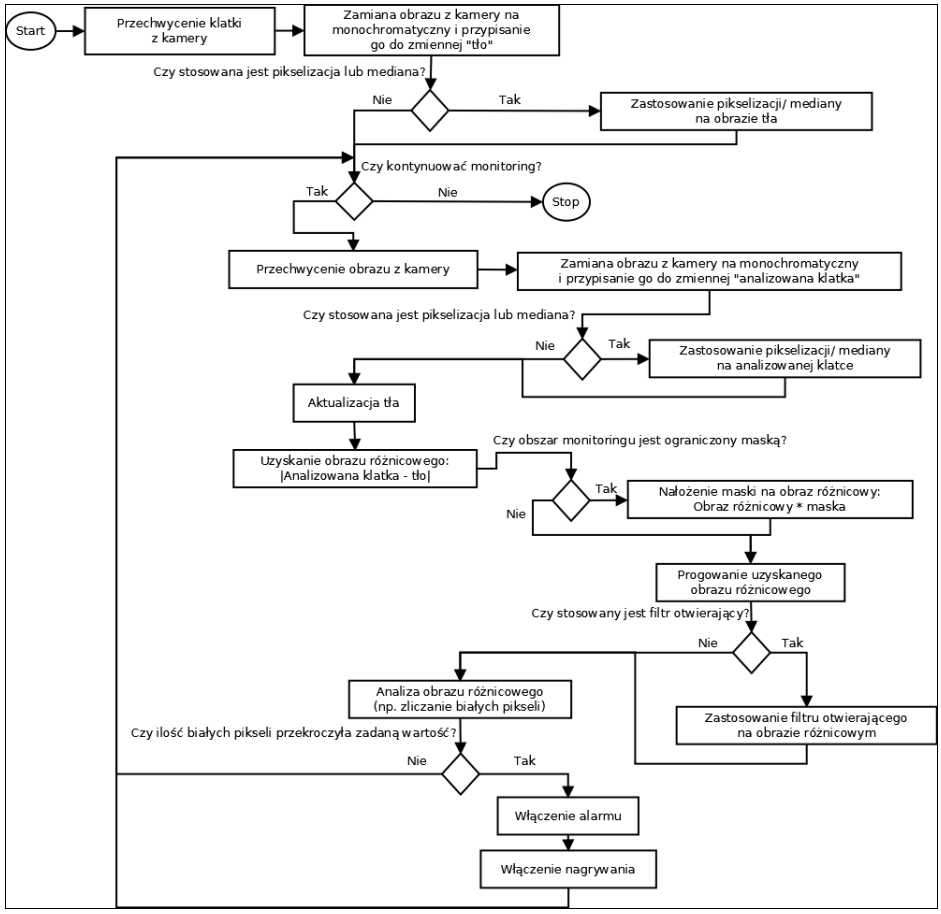
Obecnie ceny kamer przemysłowych zaczynają się od 200 złotych i mogą sięgać nawet kilku tysięcy złotych. Powstaje pytanie czy istnieje jakaś tańsza alternatywa. Okazuje się, że dla niektórych zastosowań można uzyskać zadowalający poziom bezpieczeństwa wykorzystując zwykłą kamerę internetową podłączoną do komputera, w którym zainstalowana jest aplikacja z zaimplementowanym algorytmem dokonującym analizy przechwytywanych klatek i alarmującym użytkownika o wykrytym ruchu.

2. Algorytm detekcji ruchu

Algorytm detekcji ruchu można zrealizować na wiele sposobów. Różnice mogą wynikać z wyboru klatek, które decydujemy się porównywać, wyboru metody usuwania zakłóceń, czy też metod związanych z analizą obrazu różnicowego. Algorytm operuje na czarno-białej kopii przechwyconego obrazu i w żaden sposób nie ingeruje w obraz oryginalny (oprócz ewentualnej ramki dodanej do obrazu, która może otaczać ruchomy obiekt w celu wyraźnego oznaczenia potencjalnego „intruza"). Jedną z najbardziej typowych metod jest porównanie bieżącej klatki z poprzednią. Jest to użyteczne w kompresji wideo, gdzie istnieje potrzeba przewidywania zmian w obrazie i zapisania tylko tych zmian, nie całej klatki [2]. Metoda ta, mimo że wydaje się najbardziej oczywista, nie sprawdza się najlepiej w systemach monitoringu. Różnice pomiędzy rejestrowanym klatkami mogą być bowiem tak znikome, że algorytm może w ogóle nie zadziałać.

Najbardziej efektywne algorytmy bazują na budowaniu tzw. tła sceny i porównywaniu bieżącej klatki z tłem [2]. Tło sceny uzyskuje się poprzez stałą aktualizację polegającą na nakładaniu na poprzednio uzyskane tło, bieżącej klatki. Można to zrobić za pomocą morphingu i podobnych algorytmów. Zabieg ten, przy doborze odpowiednich parametrów, dodatkowo pozwala pozbyć się problemu wykrywania regularnych ruchów (np. delikatnego ruchu liści na wietrze). Od bieżącej klatki odejmuje się tło, piksel po pikselu (co do wartości bezwzględnej) i uzyskuje obraz różnicowy. Obraz ten poddaje się następnie progowaniu, które binaryzuje obraz zamieniając piksele, których jasność przekracza zadaną wartość, na piksele białe a pozostałe na piksele czarne. Pozwala to pozbyć się już części zakłóceń. Na uzyskany obraz różnicowy nakłada się często maskę,

która ogranicza obszar monitoringu. W końcu tak przetworzony obraz może zostać poddany analizie. Najprostszą metodą jest zliczanie białych pikseli. Jeżeli ich liczba przekroczy zadaną wartość, system reaguje stanem alarmowym. Istnieją też bardziej wyrafinowane metody np. rozpoznawanie kształtów, określanie kierunku, w którym przemieszcza się obiekt, klasyfikacja ruchu.



Rys. 1. Przykładowy algorytm detekcji ruchu

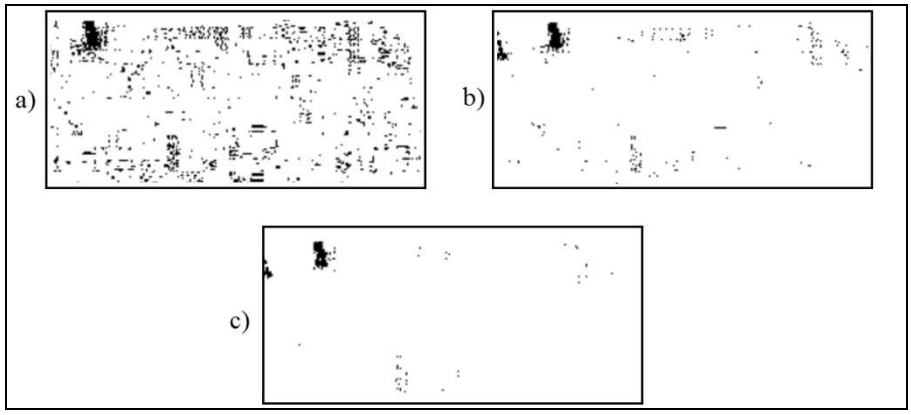
Na Rys. 1 zaznaczono miejsca, w których wykorzystuje się metody usuwania zakłóceń. Łatwo zauważyć, że pikselizacja i filtr medianowy operują na obrazach tła i bieżącej klatki, filtr otwierający wykorzystuje się zaś bezpośrednio na obrazie różnicowym po progowaniu. Metody te, w połączeniu z odpowiednim stopniem progowania, pozwalają się pozbyć zakłóceń, ale w szerszym kontekście- oprócz szumu wynikającego z niedoskonałości kamery, można pozbyć się innych niepotrzebnych informacji np.

ruchu bardzo małych obiektów. Można także sparametryzować np. ilość białych pikseli, bądź rozmiar obszarów białych pikseli wykrywanych na obrazie różnicowym, które mają powodować stan alarmu i połączyć wymienione powyżej parametry w stopień dokładności z jaką ma działać monitoring. Jak widać system zapewnia część funkcjonalności, którą można uzyskać dzięki kamerom przemysłowym i specjalizowanemu oprogramowaniu. W dalszej części artykułu chciałbym opisać wybrane przeze mnie algorytmy filtracji i dokonać ich porównania.

3. Porównanie sposobów działania wybranych metod filtracji

3.1. Progowanie

Progowanie jest kluczowym punktem algorytmu wykrycia ruchu. Operacji tej poddaje się obraz różnicowy. Progowaniu można poddać obrazy monochromatyczne. Dzięki temu otrzymujemy obraz zbinaryzowany (składający się tylko z białych i czarnych pikseli). Progowanie charakteryzuje się poziomem progowania. Jest to najczęściej wartość od 0 do 255 określająca stopień jasności piksela w przetwarzanym obrazie. Jeżeli wartość przekroczy zadany próg jasności, odpowiedniemu pikselowi przypisywany jest kolor biały, w innym wypadku piksel staje się czarny. W zależności od stosowanego algorytmu filtracji stosuje się różne poziomy progowania, im algorytm bardziej ingeruje w obraz różnicowy, tym mniejsze progowanie. Metoda ta służy jedynie do wstępnego przetwarzania obrazu, z racji że ciężko byłoby usunąć zakłócenia samym progowaniem bez utraty skuteczności wykrycia ruchu, dlatego stosuje się ją w parze z filtrami, które są względem niej komplementarne.



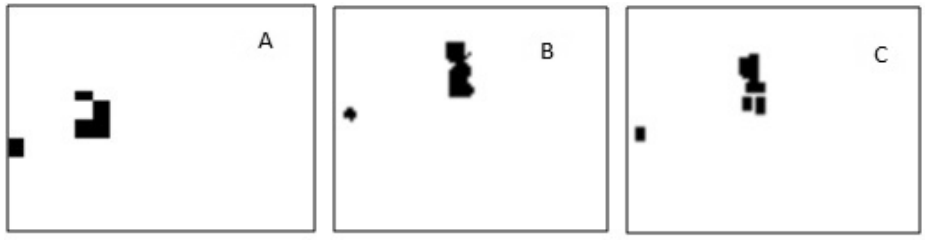
Rys. 2. Wpływ progowania na obraz różnicowy (na potrzeby wydruku odwrócono kolory): a) próg = 1; b) próg = 5; c) próg = 8

3.2. Pikselizacja

Metoda polega na wstawieniu wartości średniej pikseli z zadanego obszaru (np. 8 x 8) w miejsce oryginalnych pikseli. Działa podobnie jak filtr dolnoprzepustowy uśredniający (gdzie w masce są tylko wartości 1), obejmuje jednak większe obszary pikseli. Jest to zabieg dość radykalny dla obrazu różnicowego, dlatego stosuje się go z niskim poziomem progowania. Pikselizacji używa się na obrazie tła i przechwyconej klatce. Przykładowy wynik pikselizacji ilustruje Rys. 3A.

3.3. Filtr statystyczny medianowy

Metoda polega na posortowaniu pikseli w masce (3 x 3) rosnąco, lub malejąco i przypisanie wartości środkowej do piksela w środku maski. Filtr medianowy jest bardzo dokładny i nie ingeruje mocno w obraz różnicowy, wymaga zatem wyższego progu niż pozostałe przeze mnie opisane. Metoda skutecznie pozwala się pozbyć z obrazu zakłóceń typu pieprz i sól zachowując niemal w pełni jakość przetwarzanego obrazu. Podobnie jak pikselizację filtr medianowy stosuje się na obrazie tła i przechwyconej klatce. Przykładowy rezultat obraz różnicowy powstały w wyniku pikselizacji przedstawiono na Rys. 3B.



Rys. 3. A - obraz różnicowy powstały w wyniku pikselizacji (odwrócone kolory); B - obraz różnicowy powstały w wyniku zastosowania filtru medianowego (odwrócone kolory); C - obraz różnicowy powstały w wyniku zastosowania filtru otwierającego (odwrócone kolory)

3.4. Filtr otwierający

Filtr ten jest w rzeczywistości złożeniem dwóch innych filtrów: erozji i dylatacji. W przeciwieństwie do pozostałych opisanych przeze mnie metod usuwania zakłóceń, metoda ta nie ingeruje w obraz tła i bieżącej klatki. Operuje bezpośrednio na obrazie różnicowym po progowaniu. Najpierw przeprowadzona jest erozja (filtr minimalny), która przypisuje pikselowi wartość minimalną sąsiadujących pikseli. Pozwala to na pozbycie się z obrazu różnicowego małych grup pikseli. Erozja narusza jednak także większe obiekty, dlatego obraz poddaje się następnie dylatacji (filtr maksymalny), która przypisuje pikselowi wartość maksymalną sąsiadujących pikseli. Pozwala to przywrócić rozmiar dużym obiektom. Małe obiekty, które zniknęły z obrazu różnicowego w wyniku

erozji, nie zostaną odtworzone. Filtr ten ingeruje w obraz różnicowy słabiej niż pikselizacja, ale mocniej niż filtr medianowy. Nie wymaga stosowania wysokiego progowania. Przykładowy obraz różnicowy powstały w wyniku zastosowania filtru otwierającego przedstawiono na Rys. 3B.

4. Porównanie efektywności wybranych metod filtracji

Badając skuteczność algorytmów starałem się stworzyć warunki sprzyjające i mniej sprzyjające dla metod filtracji. Za sprzyjające warunki przyjąłem dużą różnice poziomu jasności poruszającego się obiektu w stosunku do jasności jego otoczenia. Dobrym rozwiązaniem okazało się zastosowanie wskaźnika laserowego na powierzchni białej ściany w ciemnym i doświetlonym pomieszczeniu. Badanie powtórzono dla 3 kamer różnej klasy w odległościach 2-10 metrów od ściany (wskaźnik umieszczony był zawsze w odległości 10 metrów - zmieniała się tylko odległość kamery). Wskaźnikiem świecono w losowy punkt na białej ścianie (w obrębie pola widzenia kamery). Badanie powtórzono po 10 razy w ciemnym i oświetlonym pomieszczeniu odnotowując czy nastąpiło wykrycie ruchu. Każda kamera została odpowiednio skalibrowana, a wartość progowania dobrana odpowiednio do algorytmu (była stopniowo zwiększana aż do momentu kiedy nie występowały żadne zakłócenia). Do testów użyto napisanej przeze mnie aplikacji testującej algorytmy utworzonej w języku C# z wykorzystaniem biblioteki AForge.NET i środowiska Visual Studio 2010. Aplikacja umożliwiła dynamiczny wybór metody filtracji, wybór kamery, ustalanie progu i stopnia aktualizacji tła. Na ekranie widoczne były: obraz różnicowy i obraz oryginalny, na którym, w przypadku wykrycia ruchu pojawiał się napis „ALARM”, a poruszający się obiekt był otaczany ramką. Wykorzystano sprzęt:

- kamera Logitech V-UAP41 (rozdzielczość 352 x 228 pikseli);
- kamera Creative VF0470 (rozdzielczość 640 x 480 pikseli);
- kamera Microsoft LifeCam HD-3000 (rozdzielczość 1280 x 720 pikseli);
- wskaźnik laserowy;
- notebook Toshiba Satellite L670-16M (procesor: Intel Core i3-330M 2,13 GHz,
- pamięć RAM: 4GB, karta graficzna: ATI Mobility Radeon HD 5650).

W wyniku przeprowadzonych badań uzyskano następujące rezultaty. Litera „O” oznacza oświetlone pomieszczenie a „C” ciemne pomieszczenie.

Tabela 1. Skuteczność wykrycia światła lasera dla kamery Logitech V-UAP41

Odległość [m]	Ilość wykryć (na 10 prób)					
	Pikselizacja		Filtr medianowy		Filtr otwierający	
	O	C	O	C	O	C
2	10	10	10	10	10	10
3	10	10	10	10	10	10
4	10	10	10	10	10	10
5	7	10	7	10	8	10
6	4	6	5	7	6	8
7	0	4	0	4	0	5
8	0	0	0	0	0	2
9	0	0	0	0	0	0
10	0	0	0	0	0	0

Tabela 2. Skuteczność wykrycia światła lasera dla kamery CreativeVF0470

Odległość [m]	Ilość wykryć (na 10 prób)					
	Pikselizacja		Filtr medianowy		Filtr otwierający	
	O	C	O	C	O	C
2	10	10	10	10	10	10
3	10	10	10	10	10	10
4	10	10	10	10	10	10
5	10	10	10	10	10	10
6	10	10	10	10	10	10
7	7	10	7	10	10	10
8	0	10	4	10	10	10
9	0	10	0	10	8	10
10	0	10	0	10	0	10

Tabela 3. Skuteczność wykrycia światła lasera dla kamery Microsoft LifeCam HD-3000

Odległość [m]	Ilość wykryć (na 10 prób)					
	Pikselizacja		Filtr medianowy		Filtr otwierający	
	O	C	O	C	O	C
2	10	10	10	10	10	10
3	10	10	10	10	10	10
4	10	10	10	10	10	10
5	10	10	10	10	10	10
6	10	10	10	10	10	10
7	8	10	10	10	10	10
8	2	10	6	10	10	10
9	0	10	2	10	9	10
10	0	10	0	10	6	10

Z powyższych wyników można wywnioskować, że pod względem skuteczności wykrycia ruchu przy użyciu badanych metod filtracji najlepszą metodą jest filtr otwierający, mniej skuteczna jest mediana, a najmniej skuteczna pikselizacja. Jednakże wyciągnięcie tylko takich wniosków mogłoby okazać się niezasadne, ponieważ należy także wziąć pod uwagę fakt, że w sprzyjających warunkach ilość wykryć w kamerach Creative VF0470 i Microsoft LifeCam HD-3000 wynosiła 10 na 10 dla każdej z badanych odległości. Filtr medianowy, pomimo dużej dokładności w przetwarzaniu, wymagał większego progowania niż filtr otwierający, co skutkowało spadkiem skuteczności. Duży wpływ na badanie miały możliwości kamery (szczególnie światłoczułość i rozdzielczość).

Metody filtracji przetestowano także pod kątem szybkości przetwarzania obrazów. Wykorzystano do tego aplikację testującą. Algorytmy sprawdzono dla różnych rozdzielczości obrazów wejściowych. Testowano całość algorytmu do wykrywania ruchu (nie tylko metody filtracji). Uzyskane czasy to wartości średnie czasów wykonania algorytmu z 1000 (dla rozdzielczości 160 x 120 pikseli i 320 x 240 pikseli) i 100 (dla pozostałych rozdzielczości) przebiegów.

Tabela 4. Średnie czasy wykonania przebiegu algorytmu w zależności od metody filtracji

Rozdzielczość [piksele]	Średni czas wykonania algorytmu [ms]		
	Pikselizacja	Filtr medianowy	Filtr otwierający
160 x 120	3	11	10
320 x 240	9	39	38
640 x 480	32	146	152
1280 x 960	125	559	603

Na podstawie uzyskanych wyników obliczono ile przebiegów na sekundę jest w stanie wykonać algorytm w zależności od metody filtracji. Jest to istotne, ponieważ kamera może przesyłać 30 klatek/s. Wartości w poniższej tabeli wskazują czy algorytm jest w stanie przetworzyć wszystkie przesyłane klatki (o danej rozdzielczości) w czasie rzeczywistym, czy też istnieje potrzeba ograniczenia się do pobierania klatek tylko co jakiś czas.

Tabela 5. Ilość klatek na sekundę, jaką jest w stanie przetworzyć algorytm w zależności od metody filtracji

Rozdzielczość [piksele]	Ilość klatek na sekundę, którą jest w stanie przetworzyć algorytm		
	Pikselizacja	Filtr medianowy	Filtr otwierający
160 x 120	333	90	100
320 x 240	111	25	26
640 x 480	31	6	6
1280 x 960	8	1	1

Z powyższych wartości wynika, że pikselizacja jest najszybszą z testowanych metod i na wykorzystanym sprzęcie jako jedyna była w stanie zapewnić przetwarzanie wszystkich klatek wysłanych przez kamerę dla rozdzielczości 640 x 480. Co ciekawe filtr medianowy i otwierający przetwarzają klatki w bardzo podobnym tempie (medianowy jest nieco szybszy). Z badań wynika też, że wraz ze zwiększeniem rozdzielczości analizowanych obrazów dwukrotnie, w pionie i poziomie, czas przetwarzania zwiększa się dla każdego algorytmu w przybliżeniu czterokrotnie, co jest dość oczywiste.

5. Wnioski

Wybór metody filtracji do usuwania zakłóceń z obrazu jest zależny od preferencji użytkownika. Jeżeli stosujemy kamerę o rozdzielczości co najmniej 640 x 480 pikseli, najlepszym rozwiązaniem wydaje się być pikselizacja, z racji stosunku skuteczności do

czasu przetwarzania klatki. Kamery internetowe są dosyć tanie i bardziej opłaca się zastosować lepszą kamerę i szybszy, nieco mniej dokładny algorytm, niż słabą kamerę i złożony algorytm. Jeżeli zależy nam na dużej czułości algorytmu, najlepszym rozwiązaniem jest stosowanie filtru otwierającego. Jeżeli z kolei chcemy operować na dokładnym obrazie różnicowym (np. do rozpoznawania kształtów, czy też innej wyspecjalizowanej analizy) najlepszy jest filtr statystyczny medianowy. W systemach monitoringu bywa tak, że „lepsze jest wrogiem dobrego", ponieważ duże zwiększenie czułości systemu na ruch, może w pewnym momencie skutkować fałszywymi alarmami. Należy o tym pamiętać zwłaszcza stosując taki monitoring na zewnątrz. Testując napisaną przeze mnie aplikację, kierując kamerę na zewnątrz, doszedłem do wniosku, że do zwykłego wykrycia ruchu najlepszą metodą usuwania zakłóceń jest pikselizacja. Generowała mniejszą ilość fałszywych alarmów niż pozostałe i wystarczająco skutecznie pozwalała wykryć ruch obiektów. Z moich badań wynika również, że za pomocą zwykłej kamery internetowej można uzyskać satysfakcjonujący, dla niektórych zastosowań poziom bezpieczeństwa. Serdecznie dziękuję Pani inż. Lidii Dmytrzak, Panu dr Waleremu Susłowowi, oraz Panu prof. dr hab. inż. Zbigniewowi Suszyńskiemu za wszelką udzieloną mi pomoc w przeprowadzeniu badań i pisaniu niniejszego artykułu.

Literatura

1. Kruegle H.: CCTV Surveillance Second Edition: Analog and Digital Video Practices and Technology, Elsevier, 2007.
2. Kirillov A.: Motion Detection Algorithms,
   www: http://www.codeproject.com/Articles/10248/Motion-Detection-Algorithms, 2007.

Streszczenie

Artykuł opisuje wybrane metody usuwania zakłóceń z obrazów z kamer monitorujących. Przedstawiono w nim przykładowy algorytm do wykrywania ruchu na podstawie klatek przechwyconych z kamery. Metody filtracji (pikselizacja, statystyczny filtr medianowy, filtr otwierający) zostały porównane pod względem czułości na ruch, oraz średniego czasu przetwarzania obrazów. Pojęcie „zakłócenia" w niniejszym artykule rozumiane jest jako informacja zbędna z punktu widzenia systemu monitoringu. Dotyczy więc nie tylko szumów wynikających z niedoskonałości kamery, ale również zjawisk takich jak ruchy liści, refleksy i tym podobne.

Abstract

The article describes selected methods for removing noise from images from surveillance cameras. It presents an example of an algorithm for motion detection based on frames captured from the camera. Filtration methods (pixelization, statistical median filter, the

filter opening) were compared in terms of sensitivity to movement, and the average time for processing images. The concept of "interference" in this paper shall be construed as superfluous information from the point of view of the monitoring system. This applies not only noise resulting from imperfections of the camera but also the movement phenomena, such as leaves, reflections, and the like.

Dariusz Sychel
Wydział Informatyki,
Zachodniopomorski Uniwersytet Technologiczny w Szczecinie
71-210 Szczecin,
Żołnierska 49

Redukcja czasu wykonania algorytmu Cannego dzięki zastosowaniu połączenia OpenMP z technologią NVIDIA CUDA

Słowa kluczowe: przetwarzanie równoległe, programowanie kart graficznych, CUDA, wykrywanie krawędzi, filtry splotowe, algorytm Cannego

1. Wstęp

Obecnie, jedną z popularnych technik umożliwiających zwiększenie wydajności obliczeń jest stosowanie wielordzeniowych procesorów. Jednak powstała dobra alternatywa dla tego rozwiązania, polegająca na wykorzystaniu programowalnych kart graficznych, które podobnie jak nowoczesne jednostki obliczeniowe CPU wykorzystują przetwarzanie równoległe.

W artykule autor skupia się na implementacji algorytmu Cannego, służącego do wykrywania krawędzi na obrazie opartego na połączeniu tych dwóch podejść. Testy przeprowadzone zostaną na maszynach o różnej specyfikacji, w celu sprawdzenia tego rozwiązania na słabszych pod względem mocy obliczeniowej kartach.

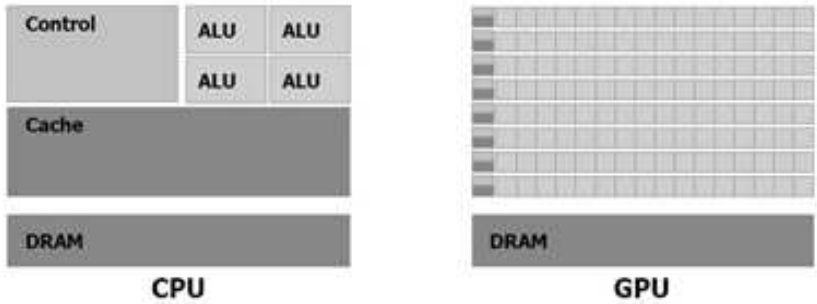
Podobne badania zostały opisane w pracach [1][2]. Podejścia tam przedstawione posiadają jednak pewne wady, zarówno w artykule [1] jak i [2] próbowano za pomocą GPU zrównoleglić algorytm oparty o BFS. Pierwsza skuteczna próba takiego rozwiązania została opisana w 2010 roku [3], 2 lata po publikacji [1]. Z artykułu tego wynika, że jest to złożone zadanie, a jego skuteczność zależy zarówno od ilości wierzchołków jak i zbalansowania grafu. W związku z tym w proponowanej implementacji zrezygnowano ze wsparcia karty graficznej dla tego kroku algorytmu na korzyść alternatywnego rozwiązania opartego o rekurencje oraz technologie OpenMP. Kolejną proponowaną zmianą jest zastąpienie stałej maski proponowanej w publikacji [2], na tworzoną dynamicznie co umożliwia sterowanie stopniem odszumienia.

W artykule opisana zostanie implementacja proponowanej wersji algorytmu oraz wyniki i analiza przeprowadzonych badań.

2. NVIDIA CUDA

Głównym zadaniem kart graficznych jest renderowanie grafiki 3D w czasie rzeczywistym. GPU (Graphic Procesor Unit) oferuje dużo większą w porównaniu do klasycznego CPU prędkość obliczeń na liczbach zmiennoprzecinkowych. Wyraźnie widać to po budowie takiego układu, w którym przeważają jednostki arytmetyczno-logiczne. Różnicę między CPU a GPU prezentuje rys. 1.

NVIDIA CUDA (Compute Unified Device Architecture) [4] jest równoległą architekturą obliczeniową stworzoną w 2006 roku przez firmę NVIDIA i do dzisiaj stosowaną w jej kartach graficznych. Architektura ta pozwala na programowanie kart graficznych w takich językach jak C, C++ czy Fortran, o ile w systemie znajdują się odpowiednie sterowniki oraz SDK. Dzięki czemu ułatwia zadanie programisty.



Rys. 1. Porównanie budowy CPU z GPU (źródło: [4])

Urządzenia zgodne z CUDA są stworzone w oparciu o architekturę SIMT (single instruction multiple thread) [4]. Każda z kart graficznych posiada macierz multiprocesorów strumieniowych, między którymi rozdzielane są wątki z bloku. Multiprocesor jest zbudowany tak, aby przetwarzał jednocześnie wiele wątków. Budowa taka pozwala na równoczesne wykonywanie znacznie większej liczby wątków niż na wielordzeniowych procesorach.

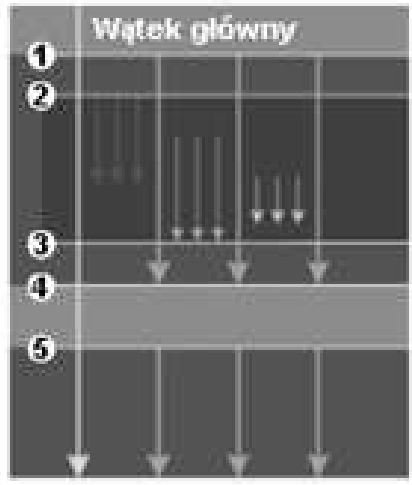
3. OpenMP

OpenMP [5] jest interfejsem programowania aplikacji na systemach z pamięcią współdzieloną, zawierających wiele procesorów. Opisuje on podział kodu, który będzie wykonywany równolegle, pomiędzy wątki na komputerach o różnej liczbie procesorów. Podobnie jak CUDA pozwala na tworzenie programów w językach: C, C++ oraz Fortran.

OpenMP w trakcie swojego działania wykorzystuje model zwany „Fork-Join" [6]. Na rys. 2 pokazane jest działanie tego modelu. Aplikacja rozpoczyna swoją pracę jako pojedynczy wątek. W momencie napotkania przez wątek bloku kodu, który ma zostać

zrównoleglony (1, 2, 5), tworzy on serie wątków pobocznych a sam zostaje wątkiem głównym.

Wątki poboczne w przypadku napotkania na zagnieżdżony blok kodu, który ma zostać zrównoleglony (2), także są w stanie utworzyć własne wątki poboczne. W momencie zakończenia wykonywania zrównoleglonego bloku kodu, wątki poboczne są kasowane, a z bloku wychodzi jedynie wątek główny dla danej sekcji (3, 4).



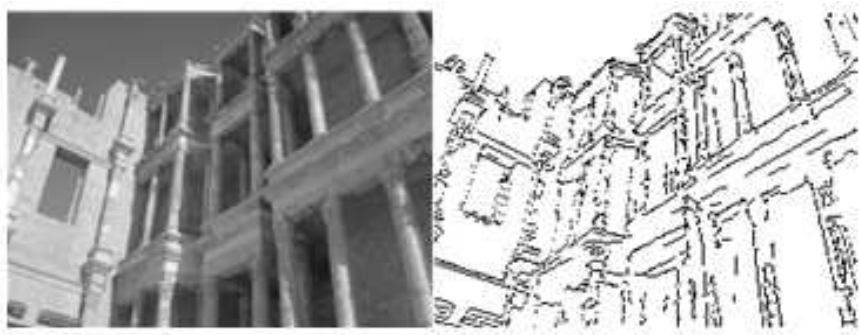
Rys. 2. Podział wątków (źródło: [6])

4. Algorytm Cannego

Algorytm Cannego jest algorytmem, którego zadaniem jest rozpoznanie krawędzi na zadanym obrazie. Metoda ta została przedstawiona przez Johna F. Cannego w roku 1986 [7]. Przykładowy efekt działania algorytmu znajduje się na rysunku 3.

Celem algorytmu jest spełnienie trzech kryteriów [8]:

- Dobra detekcja - algorytm powinien minimalizować liczbę błędów detekcji. Do błędów tego typu zalicza się zarówno błędy powstałe poprzez rozpoznanie krawędzi w miejscu, gdzie tak naprawdę nie występują oraz błędy związane z pominięciem istniejącej krawędzi.
- Dobre umiejscowienie - punkt rozpoznany jako element krawędzi powinien znajdować się jak najbliżej rzeczywistego środka krawędzi.
- Pojedyncza odpowiedź - wyznaczone krawędzie nie powinny mieć szerokości większej niż jeden piksel.



Rys. 3. Przykład działania algorytmu Cannego (źródło: własne)

Cel algorytmu można osiągnąć poprzez wykonanie czterech kroków:

1. Redukcja szumu - w tym celu można dokonać splotu obrazu z filtrem Gaussa.
2. Obliczenie natężenia oraz kierunku gradientu dla każdego punktu obrazu – w tym celu można posłużyć jednym z operatorów służących do wykrywania krawędzi (skorzystano z operatora Sobela):

$$f = \sqrt{Gx + Gy} \tag{1}$$

$$\alpha = arctg\left(\frac{Gy}{Gx}\right) \tag{2}$$

gdzie: f - macierz zawierająca natężenia gradientów, α - macierz zawierająca kierunki gradientów,

Gx, Gy - splot obrazu oryginalnego z filtrem dla kierunku x, y.

3. Usuwanie niemaksymalnych pikseli - zadaniem tego kroku jest uzyskanie krawędzi o szerokości jednego piksela, w tym celu bada się wyznaczone wcześniej kierunki oraz natężenia gradientu w sąsiadujących punktach i na tej podstawie określa się czy dany punkt jest istotny, czy musi zostać usunięty.
4. Progowanie z histerezą - ostatni krok algorytmu ma za zadanie wyeliminowanie słabych krawędzi w taki sposób, aby nie powstały luki w rozpoznanych krawędziach.

5. Implementacja

Podczas dokonywania operacji splotu obrazu, wartość każdego punktu na obrazie wynikowym jest wyznaczana niezależnie, krok 1 można zrównoleglić. Jedynym utrudnieniem jakie występuje w tej fazie algorytmu, jest samo wygenerowanie maski dla filtru (rys. 4, 1A), w związku z koniecznością obliczenia sumy elementów [9]:

$$h_g(n_1, n_2) = e^{\frac{-\left(n_1^2 + n_{21}^2\right)}{2\sigma^2}} \tag{3}$$

$$h(n_1, n_2) = \frac{h_g(n_1,n_2)}{\sum n_1 \sum n_2 h_g} \tag{4}$$

gdzie: (n1, n2) - punkt maski.

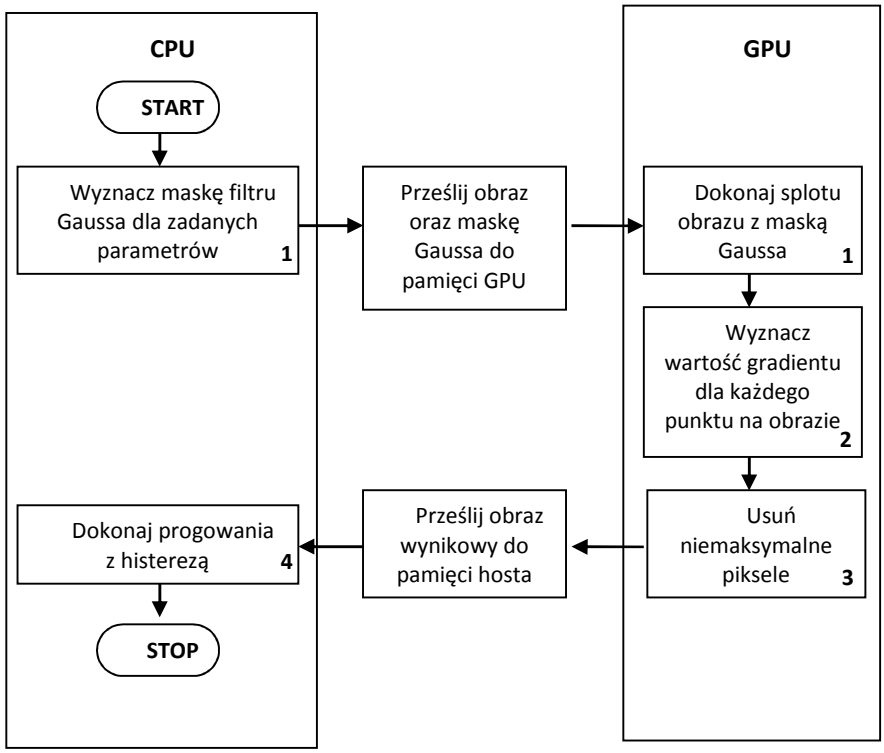
Maska taka nie może być zbyt duża, np. dla σ = 2,0 już przy wymiarach maski 10x10 utracone mogą zostać istotne krawędzie, wraz ze zwiększaniem σ efekt ten się pogarsza. Co za tym idzie nieefektywne staje się przenoszenie jej między pamięcią karty graficznej, a pamięcią hosta w celu dokonania sumowania na CPU. Dodatkowo mały rozmiar maski sprawia, że wykorzystanie OpenMP staje się nieopłacalne, dlatego część ta obliczana jest sekwencyjnie. Po wyznaczeniu wartości dla filtru maska przesyłana jest do pamięci karty i tam dokonywana jest operacja splotu. Problem ten może zostać ominięty poprzez zastosowanie stałej maski [2]. Takie podejście uniemożliwia jednak sterowanie stopniem odszumienia.

Krok 2 w związku z faktem, że zarówno natężenie gradientu jak i jego kierunek dla każdego punktu można liczyć niezależnie, o ile nie nadpisujemy oryginalnych wartości obrazu w trakcie tego procesu, także może zostać zrównoleglony. Wartości te wyznaczane są na podstawie wzorów (1) i (2).

W kroku 3, ponieważ usunięcie niemaksymalnego piksela nie wpłynie na wynik obliczeń dla sąsiednich pikseli, każdy punkt możemy rozpatrywać niezależnie. W kroku tym pobieramy dwa sąsiednie dla badanego punkty, leżące na prostej nachylonej pod wyznaczonym dla tego punktu kątem i na podstawie natężenie gradientu w tych punktach, decydujemy czy badany punkt powinien zostać zachowany. Dodatkowo w proponowanym rozwiązaniu w kroku tym przeprowadzana jest operacja progowania, krawędzią powyżej progu górnego nadawana jest wartość 255, tym które znajdują się między progiem dolnym a górnym wartość 128, pozostałe są zerowane.

Dla kroku 4 autor proponuje zmianę podejścia w stosunku do proponowanych w [1][2] poprzez rezygnacje z zastosowania dla tego kroku karty graficznej oraz zmianę zastosowanych tam algorytmów na zrównoleglony przy użyciu OpenMP algorytm rekurencyjny w celu przyspieszenia obliczeń. Proponowany algorytm, poszukuje punktów o wartości równej 255, następnie bada ich otocznie, każdemu sąsiedniemu punktowi o wartości wynoszącej 128, nadaje się wartość 255 oraz rozpoczyna się badanie jego otoczenia.

Rysunek 4 przedstawia algorytm Cannego za pomocą schematu blokowego, uwzględniając podział operacji między CPU a GPU.



Rys. 4. Zaplanowany rozkład operacji (źródło: własne)

6. Eksperymenty porównawcze

W celu określenia skuteczności badanego rozwiązania przeprowadzono dwa rodzaje badań. Kryterium pierwszego badania był czas wykonania w [ms]. Badanie polegało na porównaniu czasu wykonania czterech wersji algorytmu:

- algorytmu sekwencyjnego napisanego w języku C++ oznaczonego jako [C++],
- algorytmu zrównoleglonego przy użyciu OpenMP oznaczonego jako [OpenMP],
- algorytmu ze wsparciem przez kartę graficzną oznaczonego jako [CUDA],
- algorytmu łączącego w sobie wsparcie karty graficznej oraz zrównoleglenie poprzez zastosowanie technologii OpenMP oznaczonego jako [CUDA + OpenMP]. Kryterium drugiego badania była zgodność obrazu wygenerowanego sekwencyjnie z obrazem uzyskanym przy wsparciu GPU. Badanie to polegało na wygenerowaniu obrazu różnicowego, a następnie jego analizie w celu wykrycia czy istnieją różnice między obrazami oraz czy ewentualne różnice są istotne.

Platformy badawcze

Badania zostały przeprowadzone na sprzęcie różnej jakości. Urządzenia te różniły się między innymi liczbą rdzeni oraz mocą obliczeniową kart graficznych.

W tab. 1 znajduje się spis platform, na których testowany był algorytm w postaci tabelarycznej. Platforma 1 jest najmniej wydajnym komputerem z używanych w badaniu. Natomiast platforma 3 najbardziej, zawiera ona z 2 procesory Xeon oraz najwydajniejszą kartę graficzną z tych 3 platform.

Tab. 1. Wykaz platform badawczych

	Platforma 1	Platforma 2	Platforma 3
Procesor	Intel Core i5-480M	AMD X8 FX 8150	2 Intel Xeon E5440
Max. taktowanie	2,667GHz	3,6GHz	2,833GHz
Liczba rdzeni	2	8	4
Procesory logiczne	4	8	4
Karta graficzna	GeForce GT 540M	GeForce GTX 650	GeForce GTX 580
Wersja CC	2,1	3,0	2,0
Liczba rdzeni CUDA	96	384	512
Częstotliwość zegara	1,344GHz	1,0585GHz	1,544GHz
Max. ilość wątków	1024	1024	1024

Wyniki badań

Wyniki przedstawione w poniższych tabelach są wynikami uśrednionymi z pięciu serii testów. Tabele przedstawiają wyniki dla rozwiązania sekwencyjnego, rozwiązania opartego o OpenMP, rozwiązania wykorzystujące wsparcie GPU, oraz połączenia tych dwóch ostatnich. Dla każdej z trzech platform.

Tab. 2. Czas wykonania w [ms] dla platformy nr 1

Wielkość obrazu	C++	OpenMP	CUDA	CUDA +OpenMP
220x165	5,66	4,86	1,76	1,58
1000x1000	138,34	89,47	16,46	15,67
2000x2000	574,63	354,57	67,37	59,12
3000x3000	1253,24	1065,17	157,45	133,43
4000x4000	2838,80	1817,42	286,98	246,56

Dzięki wsparciu przez kartę graficzną dla komputera 1 przy przetwarzaniu obrazu 4000x4000px uzyskano prawie 10 krotne przyspieszenie w porównaniu z przetwarzaniem sekwencyjnym, zastosowanie dodatkowo OpenMP pozwoliło na uzyskanie prawie 12 krotnego przyspieszenia. Czas ten spadł z 2838,80ms do 246,56ms.

Tab. 3. Czas wykonania w [ms] dla platformy nr 2

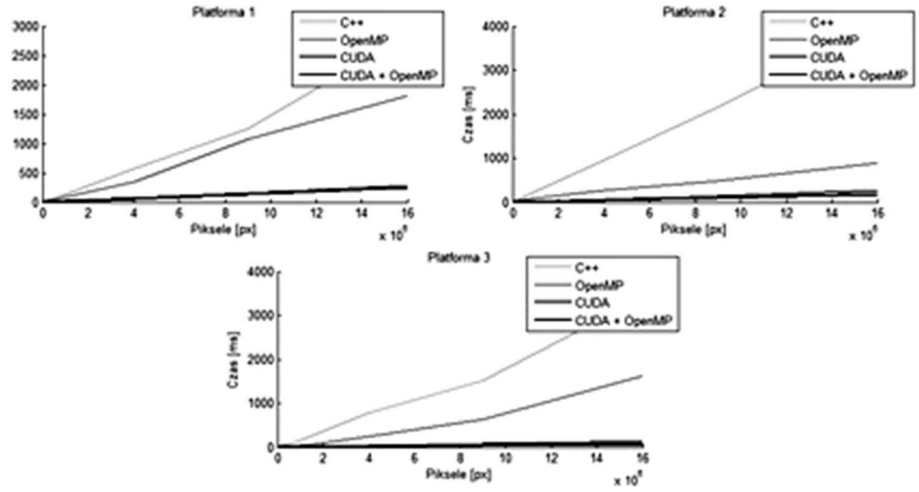
Wielkość obrazu	C++	OpenMP	CUDA	CUDA +OpenMP
220x165	8,22	4,09	1,79	1,58
1000x1000	247,17	97,56	15,55	11,13
2000x2000	963,52	262,28	60,87	46,13
3000x3000	2150,81	492,51	144,98	100,26
4000x4000	3991,20	890,84	274,91	175,60

Zastosowanie średniej klasy karty graficznej, typowej dla komputerów domowych umożliwiło dla tego obrazu uzyskanie 14,5 krotnego przyspieszenia, dzięki dodatkowemu zrównolegleniu części kodu przy użyciu OpenMP uzyskano prawie 23 krotne przyspieszenie. Z początkowych 3991,20ms spadł do 175,60ms.

Tab. 4. Czas wykonania w [ms] dla platformy nr 3

Wielkość obrazu	C++	OpenMP	CUDA	CUDA +OpenMP
220x165	10,78	3,65	1,72	1,05
1000x1000	171,92	41,58	8,20	6,41
2000x2000	775,79	256,58	31,33	20,74
3000x3000	1525,60	628,09	70,91	45,30
4000x4000	3402,15	1624,94	131,30	80,40

Zastosowanie profesjonalnej karty graficznej dla obrazu o tym samym rozmiarze spowodowało skrócenie czasu obliczeń prawie 26 krotnie w porównaniu z sekwencyjnym algorytmem wykonywanym na procesorze XEON, po dodaniu OpenMP liczba ta wzrosła do 42. Czas wykonania spadł z 3402,15ms do 80,40ms.



Rys. 5. Zestawienie wyników dla wszystkich platform

Na rys. 5 znajduje się zestawienie wykresów z wynikami dla wszystkich badanych platform.

Test różnicowy nie wykazał znaczących różnic między obrazami utworzonymi sekwencyjnie, a z wykorzystaniem karty graficznej.

7. Wnioski

Celem badania była redukcja czasu wykonania algorytmu Cannego, dzięki wykorzystaniu potencjału wielordzeniowych procesorów oraz wsparcia ze strony GPU. Dzięki takiemu rozwiązaniu udało się uzyskać nawet 42 krotne przyspieszenie w porównaniu do rozwiązania sekwencyjnego. Czas wykonania dla obrazu o rozmiarach 4000x4000 dla platformy 3 spadł z 3402,15ms do 80,40ms.

Wyniki badań pokazują, że dzięki połączeniu klasycznego podejścia ze wspomaganiem w postaci karty graficznej w celu zrównoleglenia algorytmu Cannego uzyskano znaczące skrócenie czasu obliczeń od 12 do 42 razy dla obrazu 4000x4000 w stosunku do tradycyjnego, sekwencyjnego przetwarzania. Fakt ten może zostać wykorzystany w wielu dziedzinach nauki, techniki, przemysłu, gdzie algorytmy wykrywania krawędzi mają zastosowanie, jak na przykład:

- szybsze przetwarzania obrazów medycznych (w konsekwencji szybsza diagnoza),
- analiza zdjęć satelitarnych (programy typu GIS),
- usprawnienie procesu digitalizacji dokumentów - OCR (wyższa efektywność instytucji mogących operować na dokumentach w wersji cyfrowej).

W celu uzyskania tego efektu nie jest wymagana droga karta graficzna, dlatego rozwiązanie to może mieć także zastosowanie w programach służących do przetwarzania obrazów pisanych na komputery domowe czy laptopy.

8. Literatura

1. Luo Y., Duraiswami R.: Canny edge detection on NVIDIA CUDA. 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 43, 1, 2008, 1-8
2. Ogawa K., Ito Y., Nakano K.: Efficient Canny Edge Detection Using a GPU. ICNC '10 Proceedings of the 2010 First International Conference on Networking and Computing, 2010, 279-280
3. Luo L., Wong M., Hwu M.: An effective GPU implementation of breadth-first search. Design Automation Conference (DAC), 2010 47th ACM/IEEE, 2010, 52-55 [4] NVIDIA.: NVidia CUDA C Programing Guide, Version 4.2.
4. https://www.math.umass.edu/~johnston/CUDA_WG_2012/CUDA_C_Programming_Guide.pdf, dostęp 06.04.2013
5. Blaise B.: OpenMP. https://computing.llnl.gov/tutorials/openMP/, dostęp 06.04.2013
6. Gatlin K. S., Isensee P.: Reap the Benefits of Multithreading without All the Work. http://msdn.microsoft.com/pl-pl/magazine/cc163717%28en-us%29.aspx, dostęp 06.04.2013
7. Canny F.J.: A Computational Approach to Edge Detection. J-IEEE-PAMI, 8, 6, 1986, 679-698.
8. Pratt W.K..: Digital Image Processing - PIKS Scientific Inside. John Wiley & Sons, 2007.
9. The MathWorks, Inc.: Documentation Center. http://www.mathworks.com/help/images/ref/fspecial.html, dostęp 06.04.2013

Streszczenie

Artykuł prezentuje alternatywne podejście do programowania równoległego poprzez wykorzystanie programowalnych kart graficznych w celu wsparcia obliczeń, oraz połączenie tego podejścia z klasycznym zrównolegleniem opartym o wielordzeniowe procesory. Przeprowadzone testy przedstawiają zysk czasu jaki można uzyskać dzięki odpowiedniemu połączeniu OpenMP z technologią CUDA w obliczeniach związanych z wykrywaniem krawędzi na obrazie rastrowym przy użyciu algorytmu Cannego. Badania przeprowadzone zostały na sprzęcie różnej jakości. Napisane algorytmy są zgodne z CC 1,0 (zdolność obliczeniowa karty graficznej).

Abstract

This paper presents an alternative approach to parallel programming by using programmable graphics card to support calculations and combines this approach with a classical parallelization based on multi-core processors. The tests show the gain time that can be achieved through a combination of OpenMP with CUDA technology in the calculation of the edge detection on the raster image using the Canny's algorithm. Tests were carried out on the equipment of varying quality. The algorithms are compatible with CC 1.0 (compute capability graphics card.)

Słowa kluczowe: parallel processing, programming, graphics cards, CUDA, edge detection filters, the Canny's algorithm

Piotr Ratuszniak
Łukasz Gątnicki
Wydział Elektroniki i Informatyki
Politechnika Koszalińska
ratusz@ie.tu.koszalin.pl
lukasz.gatnicki@gmail.com

Aplikacja wyszukiwania i wizualizacji trasy dla przewoźników samochodowych

Słowa kluczowe: nawigacja, GPS, optymalizacja trasy, problem komiwojażera, algorytmy genetyczne.

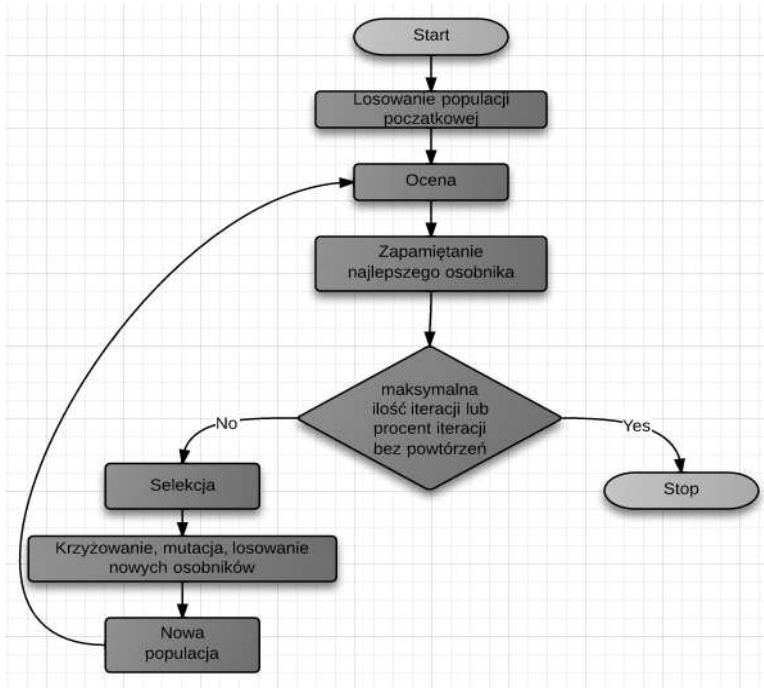
1. Wstęp

Jednym z przełomowych wydarzeń w dziedzinie transportu i nawigacji było udostępnienie w latach osiemdziesiątych przez kongres amerykański systemu GPS (ang. *Global Positioning System*). Amerykański wojskowy system został udostępniony do wykorzystania cywilnego bez jakichkolwiek opłat co zapewne miało decydujący wpływ na jego szybkie upowszechnienie. Obecnie system ten jest wykorzystywany w wielu dziedzinach życia codziennego, np.: w nawigacji samochodowej, systemach nawigacji dla pieszych i rowerzystów, w pozycjonowaniu do określenia miejsca bieżącego pobytu oraz w szeroko pojętym określaniem czasu [1]. Z drugiej strony bardzo dynamiczny rozwój urządzeń mobilnych wyposażonych w odbiorniki GPS spowodował wykorzystanie tego systemu na masową skalę. Pojawiło się wiele różnych aplikacji wykorzystujących ten system, jednak jedną z największych popularności cieszą się aplikacje do nawigacji samochodowej. Obecnie na rynku istnieje wiele aplikacji do nawigacji samochodowej, sprzedawanych często razem z odbiornikami systemu GPS. W Polsce jedną z najbardziej popularnych i najbardziej zaawansowanych aplikacji do nawigacji samochodowej jest AutoMapa [2]. W aplikacji tej mamy kilka różnych scenariuszy wyszukiwania trasy przejazdu. Pomimo ponad dziesięcioletniego pobytu na rynku aplikacja ta, podobnie jak inne popularne aplikacje dostępne na rynku, nie zawiera zaawansowanego algorytmu minimalizacji trasy przejazdu dla wielu zdefiniowanych punktów przejazdu w sposób analogiczny dla problemu komiwojażera. Opcja wyszukiwania trasy analogicznie do tego problemu będzie z pewnością miała szerokie zastosowanie w wielu firmach kurierskich i przewozowych, jak również dla wszystkich przewoźników, dla których kursy realizowane są według scenariusza „baza - wiele punktów docelowych- baza”. Brak tego rodzaju optymalizacji wyszukiwanej trasy

w popularnych aplikacjach do nawigacji samochodowej był głównym motywem powstania opisywanej aplikacji. W artykule opisano utworzoną aplikację do optymalizacji trasy przejazdu według opisanego scenariusza, bez konieczności określania kolejności punktów pośrednich, jak ma to miejsce w powszechnie dostępnych aplikacjach do nawigacji samochodowej. Utworzona aplikacja pobiera powszechnie dostępne dane na temat niezbędnych miejscowości oraz odległości pomiędzy nimi za pomocą Internetu oraz dokonuje optymalizacji trasy przejazdu poprzez odpowiedni dobór kolejności punktów pośrednich. Wynikiem działania aplikacji jest wizualizacja wyznaczonej trasy przejazdu na mapie, wygenerowanie wskazówek dojazdu w postaci listy z kolejnymi manewrami. Aplikacja posiada również możliwość wygenerowania pliku z opisem trasy, umożliwiającego jej wczytanie do popularnej aplikacji nawigującej – AutoMapa [2], co z pewnością podnosi jej walory użytkowe.

2. Algorytm genetyczny do optymalizacji trasy

Przedstawiony powyżej problem optymalizacji trasy w algorytmice znany jest pod nazwą Problemu Komiwojażera (z ang. TSP - Travelling Salesman Problem). Problem komiwojażera zaliczany jest do grupy problemów NP-trudnych [3, 4, 5] i jest to zagadnienie natury optymalizacyjnej należące do działu matematyki i informatyki zwanego teorią grafów. Problem ten dotyczy odnalezienia minimalnego cyklu Hamiltona w grafie pełnym ważonym. Innymi słowy chodzi o znalezienie ścieżki po krawędziach grafu o najmniejszej sumie wag, która wychodząc z zadanego punktu będzie przebiegać przez każdy z pozostałych wierzchołków tylko raz i wróci do wierzchołka startowego. W powyższym opisie również łatwo dostrzec podobieństwo teorii do praktycznej pracy wykonywanej przez kuriera. Do jego obowiązków należy bowiem wyruszenie z miasta bazy do wszystkich odbiorców rozlokowanych w różnych miejscowościach i powrót do punktu startu. Obecnie znanych jest wiele metod rozwiązywania problemu komiwojażera wykorzystujących np.: algorytmy genetyczne [6, 7], mrówkowe [8] i memetyczne [9]. W opisywanej aplikacji, na jej bieżącym etapie rozwoju, do wyszukiwania trasy w aplikacji został wykorzystany algorytm genetyczny. W stosunku do klasycznego algorytmu genetycznego zostały wprowadzone pewne modyfikacje. Zmiany te pozwalają na zmniejszenie zbieżności algorytmu do ekstremum lokalnego oraz zmniejszenie czasu jego działania z minimalnym wpływem na jakość wyników. Ogólny schemat działania opracowanego algorytmu przedstawiony jest na rys. 1.

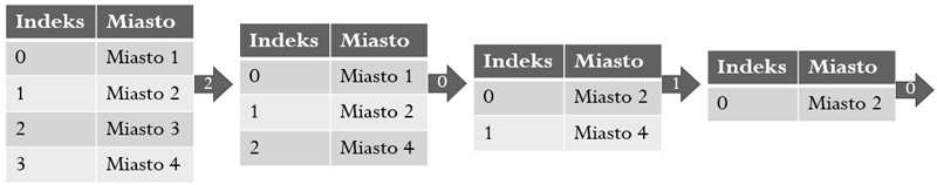


Rys. 1. Schemat blokowy zaimplementowanego algorytmu genetycznego

2.1. Reprezentacja osobników

Jednym z najważniejszych etapów tworzenia algorytmu genetycznego jest określenie reprezentacji danych. W aplikacji wyszukiwania i wizualizacji trasy ważne było znalezienie reprezentacji, która nie będzie znacząco komplikować losowania populacji oraz działania operatorów genetycznych krzyżowania i mutacji. Reprezentacja w postaci prostej listy z nazwami miast mogłaby powodować generowanie osobników niespełniających ograniczeń trasy, w taki sposób, że po operacji krzyżowania i mutacji należałoby sprawdzić poprawność powstałych osobników (tras) w celu wykluczenia zduplikowanych miast. Istnieje co prawda możliwość stosowania mechanizmów naprawczych [10] po zastosowaniu operatorów algorytmu genetycznego, jednak stosowanie tego typu mechanizmów powoduje zwiększenie złożoności obliczeniowej każdej generacji algorytmu, co w konsekwencji może powodować znaczny spadek wydajności całego algorytmu. Z tego powodu ważnym zagadnieniem jest odpowiedni dobór reprezentacji danych. W opisywanej aplikacji zastosowano reprezentację danych w postaci listy odniesień [10]. Polega ona na utworzeniu listy odniesień do listy wszystkich miast w kolejności ich pobierania. Punktem odniesienia w tej reprezentacji jest wprowadzona przez użytkownika do programu lista wszystkich miast, które należy odwiedzić. Każdy osobnik jest odzwierciedleniem tego, w jakiej kolejności są

odwiedzane miasta z głównej listy. Po dopisaniu do listy referencji kolejnej wartości z listy głównej usuwany jest jej odpowiednik. W wyniku tego na *i*-tej pozycji dla $i = 0..n - 1$, gdzie: n - liczba miast do odwiedzenia, znajdzie się zawsze liczba całkowita z przedziału $[0, n - 1 - i]$. Tworzenie przykładowej listy odniesień przedstawiono na rys. 2.



Lista odniesień	Miasta
2	Miasto 3
0	Miasto 1
1	Miasto 4
0	Miasto 2

Rys. 2. Przykładowe kroki w tworzeniu listy odniesień

Taka reprezentacja sprawia jednak pewne trudności w opracowaniu funkcji celu oceny. Aby możliwe było wyliczenie długości tras do ocenienia rozwiązań niezbędne jest przejście do standardowej reprezentacji w postaci listy miast w kolejności odwiedzania. Wiąże się to z koniecznością wykonania dodatkowych operacji i zwiększenia złożoności obliczeniowej funkcji celu, jednak ta dodatkowa złożoność obliczeniowa jest rekompensowana podczas wykonywania operacji krzyżowania i mutacji. Biorąc pod uwagę wzór na *i*-ty element listy referencji, zarówno podczas operacji standardowego jednopunktowego krzyżowania oraz standardowej operacji mutacji wybranej pozycji, za każdym razem generowany jest nowy osobnik reprezentujący dopuszczalne rozwiązanie problemu. Podczas programowej implementacji w aplikacji do przechowywania informacji o osobniku została zaprojektowana dodatkowa pomocnicza struktura. Poza tablicą zawierającą referencje w postaci indeksów do listy miast dodatkowa struktura posiada tablicę zawierającą odległości cząstkowe trasy oraz zmienną przechowującą całkowitą jej długość.

2.2. Funkcja oceny i operatory algorytmu

Operatory algorytmu

Ponieważ aplikacja wyszukiwania i wizualizacji trasy oparta jest na problemie komiwojażera w zaimplementowanym algorytmie genetycznym wartością funkcji oceny każdego osobnika jest długość trasy, którą on reprezentuje. Po eksperymentalnych doświadczeniach zaobserwowano fakt, że trasy reprezentowane przez poszczególnych osobników różnią w sposób znaczący. Aby wyeliminować możliwość powstawania superosobników, co oznaczałoby zwiększenie prawdopodobieństwa utknięcia algorytmu w minimum lokalnym, zastosowano selekcję metodą rankingu liniowego [11]. Po wyliczeniu odległości całkowitej każdego osobnika cała populacja jest sortowana malejąco. Wartość prawdopodobieństwa przejścia do kolejnej generacji jest wyliczana na podstawie ilorazu położenia osobnika w posortowanej liście przez sumę położeń wszystkich osobników. Przykładowo mając trzy osobniki posortowane malejąco 3,2,1 pierwszy osobnik na liście będzie posiadał prawdopodobieństwo $\frac{1}{6}$, drugi $\frac{2}{6}$, trzeci $\frac{3}{6}$.

Po operacji selekcji, w kolejnym kroku algorytmu, wykonywane są operatory krzyżowania, mutacji oraz wszczepiania do populacji nowych losowych osobników. Po szeregu doświadczeń eksperymentalnych, realizowanych dla założonych zestawów punktów pośrednich trasy, określono następujące parametry wymienionych operatorów:

- Krzyżowanie – dzieli osobniki rodzicielskie w okolicach połowy trasy na dwie części i zamienia je krzyżowo. Operacja ta zachodzi z prawdopodobieństwem 50%.
- Mutacja – maksymalnie trzykrotnie wybierana jest losowa wartość genu osobnika i następnie zamieniana ją z inną losową wartością z odpowiedniego przedziału. Operacja zachodzi każdorazowo z prawdopodobieństwem 20%.
- Zamiana osobnika na losowego – w miejsce osobnika rodzicielskiego generowany w sposób losowy nowy osobnik. Operacja zachodzi z prawdopodobieństwem 30%.

Operacja zamiana osobnika na losowego nie jest standardowym operatorem genetycznym, jednak została wykorzystana w celu zmniejszenia zbieżności algorytmu. Losowa zmiana na zupełnie nowego osobnika wyklucza w jeszcze większej mierze możliwość powstania superosobników i utknięcia algorytmu w minimum lokalnym. Operacja ta miała widoczny wpływ na jakość uzyskiwanych rozwiązań.

2.3. Warunki zakończenia pracy algorytmu

Z przedstawionego ogólnego schematu blokowego działania opracowanego algorytmu genetycznego można wywnioskować, że zakończenie pracy algorytmy genetycznego następowało po obliczeniu założonych, eksperymentalnie ustalonych dla

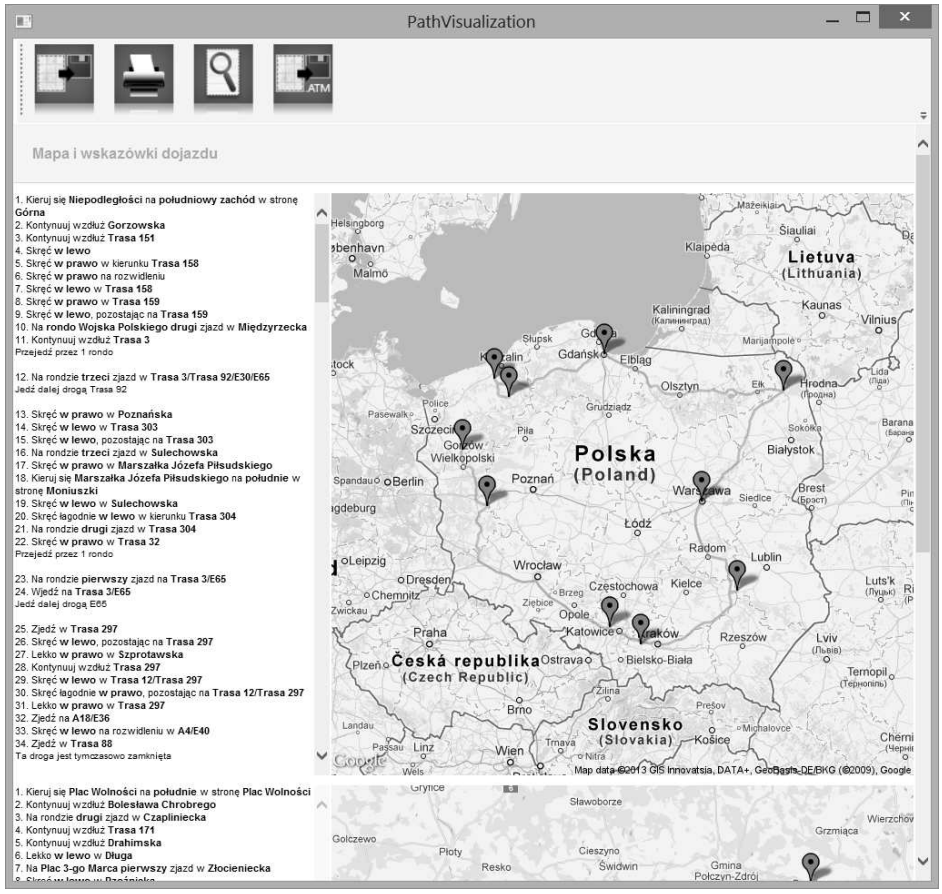
danego przedziału punktów pośrednich trasy obliczonych generacji lub po uzyskaniu procentowej wartości krytycznej bez poprawy jakości uzyskanego rozwiązania.

3. Aplikacja wyszukiwania i wizualizacji trasy

Aplikacja wyszukiwania i wizualizacji trasy dla przewoźników samochodowych została wykonana w celu usprawnienia procesu planowania tras. Umożliwia ona planistom lub samym przewoźnikom odnalezienie suboptymalnej trasy przejazdu do zadanych miejscowości, na podstawie ich nieuporządkowanej listy. Na rysunku 3 przedstawiono okno główne aplikacji z listą dodanych punktów trasy.

Rys. 3. Okno główne aplikacji z listą punktów trasy

Aplikacja pozwala także wygenerować i wydrukować mapy wraz ze wskazówkami dojazdu oraz plik do urządzenia nawigacji samochodowej korzystającej z oprogramowania AutoMapa. Program umożliwia modyfikacja parametrów algorytmu genetycznego użytego do wyszukiwania trasy oraz zapis map i wskazówek dojazdu w formie pliku HTML. Na rys. 4 przedstawiono okno aplikacji z wygenerowaną mapą z zaznaczoną zoptymalizowaną trasą przejazdu oraz z wygenerowanymi tekstowymi wskazówkami dojazdu.



Rys. 4. Wygenerowana mapa z zaznaczoną trasą przejazdu oraz wygenerowane tekstowe wskazówki dojazdu

3.1. Wykorzystane narzędzia i technologie

Aplikację wyszukiwania i wizualizacji trasy dla przewoźników samochodowych została wykonana w oparciu o technologie firm Microsoft i Google. Platformą, z której skorzystano podczas tworzenia programu jest .NET Framework i język C#. Ponadto podczas realizacji aplikacji wykorzystano następujące narzędzia i technologie:

- Windows Presentation Foundation (WPF) – jako nowoczesny silnik graficzny i API pozwalające na budowanie interfejsu aplikacji korzystając ze znacznikowego języka XAML opartego na formacie XML;

- Task Parallel Library (TPL) – zestaw bibliotek ułatwiający wprowadzenie do aplikacji elementów przetwarzania równoległego i współbieżności, wykorzystanych w mechanizmie pobierania danych o miejscowościach pośrednich trasy;
- ADO.NET – zbiór bibliotek zapewniających dostęp do baz danych;
- Framework MVVM Light - zestaw narzędzi i komponentów ułatwiający i przyspieszający tworzenie aplikacji zgodnych z wzorcem MVVM w technologiach WPF, Silverlight i Windows Phone. Rozwiązania dostępne w tym zestawie pomagają w zachowaniu separacji kodu pomiędzy klasami modelu i widoku;
- serwer bazodanowy Microsoft SQL Server 2012 w darmowej wersji Express do przechowywania danych na temat miejscowości;
- serwisu Google Maps - do dostarczenia danych o odległościach, map i wskazówek dojazdu za pomocą ogólnodostępnego API w postaci Web Services. Pobranie danych z tego serwisu było możliwe poprzez budowanie odpowiednich adresów URL.

3.2. Funkcjonalności aplikacji

Aplikacja składa się z dwóch zasadniczych modułów: moduł odpowiedzialny za wyszukiwanie trasy oraz moduł służący do jej wizualizacji. Dla obu modułów aplikacji poniżej zamieszczono listy ich wybranych funkcjonalności.
Wybrane funkcjonalności, możliwości i ograniczenia modułu wyszukiwania trasy:

- możliwość wykorzystania algorytmu wyszukiwania trasy dla liczby miast mieszczącej się w przedziale od 5 do 99;
- wykorzystanie przez algorytm wyszukiwania trasy bazy danych odległości opartej na informacjach pozyskanych z serwisu Google Maps;
- autouzupełnianie pola tekstowego używanego do wprowadzania nazw miejscowości na podstawie listy miast znajdującej się w bazie danych;
- wyświetlanie w formie listy w głównym oknie programu miast ułożonych w odpowiedniej kolejności po zakończeniu działania algorytmu genetycznego;
- możliwość skorzystania z predefiniowanych parametrów algorytmu wyszukiwania takich jak liczebność populacji, liczba iteracji i współczynnik iteracji bez zmian;
- możliwość definiowania własnych ustawień algorytmu wyszukiwania, takich jak: liczebność populacji, liczba iteracji i współczynnik iteracji bez zmian;

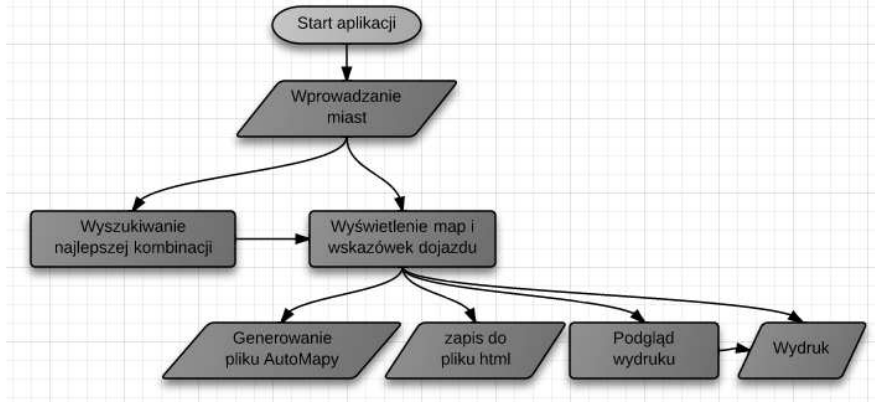
- informacja o działaniu algorytmu wyszukiwania trasy za pomocą paska postępu pracującego w trybie niekreślonym, w pasku statusu w głównym oknie aplikacji;
- możliwość skorzystania z pomocy dostępnej w aplikacji dotyczącej ustawień i parametrów algorytmu wyszukiwania trasy.

Wybrane funkcjonalności, możliwości i ograniczenia modułu wizualizacji trasy:

- przedstawienie tras za pomocą map z zaznaczoną wyraźną linią i tekstowych wskazówek dojazdu w oknie dialogowym aplikacji za pomocą usług webowych Google Maps;
- przedstawienie na jednej mapie maksymalnie dziesięciu punktów trasy-wskazówki dojazdu osobne dla każdej mapy;
- mapy i wskazówki dojazdu przedstawione oknie dialogowym w formie stosu jedna pod drugą;
- możliwość wydruku map wraz ze wskazówkami dojazdu z możliwością podglądu wydruku;
- możliwość wygenerowania pliku ATM pozwalającego na przeniesienie trasy do odbiornika GPS z oprogramowaniem AutoMapa;
- możliwość zapisu map wraz ze wskazówkami dojazdu do pliku HTML;
- możliwość skorzystania z pomocy dostępnej w aplikacji, dotyczącej korzystania z modułu wizualizacji trasy.

3.3. Schemat działania aplikacji

Głównym celem zaprojektowanej aplikacji jest wyszukiwanie najlepszej możliwej trasy ułożonej na podstawie wprowadzonej przez użytkownika listy, a następnie przedstawienie jej za pomocą graficznych map i tekstowych wskazówek. Nie jest to jednak jedyny dostępny scenariusz wykorzystania, co obrazuje rysunek 5.



Rys. 5. Ogólny schemat działania aplikacji

Aplikacja, poza wyżej wspomnianym scenariuszem, może służyć do pobrania trasy dla dowolnie wprowadzonych i ułożonych przez użytkownika punktów. Nie jest konieczne uruchamianie algorytmu wyszukiwania najlepszej kombinacji. Poza tym po pobraniu i wyświetleniu map i wskazówek istnieje możliwość ich wydruku, zapisu do pliku *.html, możliwego do otwarcia w dowolnej przeglądarce, a także wygenerowania pliku *.atm współpracującego z urządzeniami wyposażonymi w odbiornik GPS z oprogramowaniem AutoMapa.

4. Podsumowanie

Zaprojektowana aplikacja jest potrzebnym rozszerzeniem istniejącego oprogramowania systemów GPS poprzez dodanie możliwości wyszukania drogi na podstawie wielu punktów bez określania ich kolejności. Aplikacja za pomocą algorytmu genetycznego wyznacza kolejność miejscowości wpisanych przez użytkownika. Uporządkowana w ten sposób lista to rozwiązanie zapewne suboptymalne, jednak uzyskane w rozsądnym czasie. Wbudowany moduł wyświetlania tras i wskazówek dojazdu oraz możliwość ich drukowania pozwalają na korzystanie z nich w podróży. Gdyby jednak ich statyczna forma okazała się niewystarczająca dla użytkownika można skorzystać z wyznaczonej trasy na odbiornikach GPS poprzez wygenerowanie odpowiedniego pliku. W obecnej wersji wspierane jest oprogramowanie jednego z najbardziej popularnych tego typu rozwiązań - AutoMapa. Dzięki tym możliwościom aplikacja jest oprogramowaniem dostosowanym do potrzeb przewoźników samochodowych i zapewne znajdzie zastosowanie, gdyż wypełnia dość istotną lukę na rynku aplikacji nawigacji samochodowej.

Zaprezentowana aplikacja cechuje się dużymi możliwościami rozbudowy. Możliwe jest rozszerzenie bazy danych odległości o kolejne miejscowości niebędące miastami Polski i dzięki temu możliwe jest uwzględnienie obsługi innych państw Europy. Ponadto planowane jest dalsze usprawnianie działania samego algorytmu genetycznego poprzez jego równoległą realizację, np. z wykorzystanie powszechnie już dostępnych procesorów wielordzeniowych. Równoległa implementacja zapewne pozwoli jeszcze zredukować czas optymalizacji trasy. Usprawnienia mogą dotyczyć również interfejsu użytkownika lub obsługi większej ilości systemów GPS. Ciekawą opcją może być też stworzenie wersji działającej na urządzeniach mobilnych z wbudowanymi modułami GPS.

Bibliografia

1. http://www.gps.gov/applications/
2. http://www.automapa.pl/?PEI=35489&lng=PL
3. Papadimitriou, Christos H. (1977), "The Euclidean traveling salesman problem is NP-complete", Theoretical Computer Science 4 (3): 237–244,

4. Applegate, D. L.; Bixby, R. M.; Chvátal, V.; Cook, W. J. , The Traveling Salesman Problem, 2006, ISBN 0-691-12993-2
5. Gutin, G.; Punnen, A. P. (2006), The Traveling Salesman Problem and Its Variations, Springer, ISBN 0-387-44459-9
6. F. H. Khan, N. Khan, S. Inayatulla, S. Nizami: Solving ISP Problem by Using Genetic Algorithm, International Journal of Basic & Applied Sciences IJBAS-IJENS Vol:09 No:10 ,2009
7. M.Karova, V.Smarkov, S. Penev: Genetic operators crossover and mutation in solving the TSP problem, International Conference on Computer Systems and Technologies, 2005
8. Marco Dorigo, Luca Maria Gambardella: Ant colonies for the travelling salesman problem, Biosystems, Volume 43, Issue 2, Elsvier, July 1997, Pages 73–81
9. Peter Merz, Bernd Freisleben: Memetic Algorithms for the Traveling Salesman Problem, Complex Systems, 13 (2001) 297–345; 2001
10. Z. Michalewicz, „Algorytmy genetyczne + struktury danych = programy ewolucyjne", Wydawnictwo Naukowo-Techniczne, 1996
11. David E. Goldberg, Kalyanmoy DebA comparative analysis of selection schemes used in genetic algorithms, Foundations of Genetic Algorithms, 1991

Streszczenie

W artykule zaprezentowano praktyczną implementację algorytmu genetycznego do rozwiązywania problemu optymalizacji trasy analogicznego do problemu komiwojażera. Algorytm został zaimplementowany w autorskiej aplikacji do wyznaczania trasy przejazdu dla rzeczywistych danych geograficznych polskich miejscowości pobieranych z serwisu Google Maps. Prezentowana aplikacja generuje wskazówki dojazdu i umozliwia export wyznaczonej trasy do programu Automapa, co stanowi jego doskonałe uzupełnienie.

Abstract

The paper presents a practical implementation of a genetic algorithm to solve the problem of route optimization analogous to the traveling salesman problem. The algorithm has been implemented in the author's application for route calculation for the real Polish geographic data retrieved from Google Maps service. Presented application generates travel directions in the text and graphic form and allows to export the computed route to the Automapa program, which is his perfect complement.

Słowa kluczowe: vehicle navigation system, GPS, route optimization, travelling salesman problem, genetic algorithms.

Przemysław Plecka
Krzysztof Bzdyra
Wydział Elektroniki i Informatyki
Politechnika Koszalińska
Ul. Śniadeckich 2, 75-453 Koszalin
tel.: +48602336363,

Wybór metod szacowania kosztów modyfikacji na wstępnych etapach cyklu życia oprogramowania ERP

Keywords: ERP, implementation, cost estimation

Wstęp

W chwili obecnej każdy z liczących się na rynku producentów posiada w swoim portfelu produkt standardowy klasy ERP: SAP - Business Suite, Microsoft - Dynamix AX, JD Edwards - EnterpriseOne, itp. Podczas rozmów handlowych w trakcie sprzedaży systemów strony dochodzą jednakże do wniosku, że organizacja procesów w przedsiębiorstwie nie pokrywa się całkowicie z procesami realizowanymi przez oferowany system informatyczny [1]. Istnieje grupa procesów, której nie odpowiada żadna funkcjonalność w standardowym systemie ERP. Na tym etapie pojawia się potrzeba przystosowania systemu informatycznego (skr. SI) do przedsiębiorstwa. Koszty modyfikacji systemu podnoszą wartość kontraktu. W niektórych przypadkach alternatywą jest przystosowanie przedsiębiorstwa do systemu informatycznego, jednakże koszty zmiany organizacji obciążą wówczas klienta. Dopiero gdy klient pozna koszt, jaki będzie musiał ponieść na wdrożenie systemu (w tym przystosowanie), skłonny jest do rozważenia zmian w swojej organizacji. Aby dostawcy mieli podstawy do negocjacji, istotnym jest szacowanie kosztów na jak najwcześniejszym etapie.

Prowadząc proces wyliczania kosztów modyfikacji dostawcy napotykają na trudności w doborze odpowiedniej metody. Zwykle na początkowym etapie wybierają jedną, najlepiej im znaną metodę i stosują ją przez cały okres wycen. Taka praktyka doprowadza do większych błędów szacowań [2]. Pomocne dla dostawców byłoby narzędzie sugerujące metodę za pomocą, której uzyskane zostanie najdokładniejsze szacowania. Po każdym etapie procesu sprzedaży dostawca mógłby zweryfikować dane jakie zebrał i uzyskać sugestie jakie metody szacowania zastosować. Wiadome są więc etapy cyklu życia oprogramowania [3] oraz metody szacowania oprogramowania. Poszukiwana jest odpowiedź na pytanie, jaka jest procedura wyboru metod szacowania oprogramowania na danym etapie fazy strategicznej procesu wdrożenia SI. Zakres problemu ograniczony

został do systemów informatycznych klasy ERP wdrażanych w średnich przedsiębiorstwach produkcyjnych.

Metody pomagające wycenić koszty wykonania oprogramowania są znane i opisane w literaturze, np. przez McConella [4]. Jednakże z powodu zmian technologii informatycznych powszechność tych metod ciągle ulega zmianie. Zastosowanie algorytmicznych metod w początkowych etapach projektów informatycznych jest trudne. Nie istnieją wówczas jeszcze dokumenty projektowe zawierające dane potrzebne algorytmom estymującym. Mimo że przykłady zastosowania metod algorytmicznych we wczesnych etapach projektów informatycznych można znaleźć w literaturze[5] [6], to praktyka dostawców systemów informatycznych pokazuje stosowanie metod niealgorytmicznych jako szybszych (tzn. tańszych) i łatwiejszych do realizacji. W literaturze możemy znaleźć różne przykłady sugestii zastosowania metod szacowania kosztów dla projektów informatycznych, począwszy od stwierdzeń, że należy stosować dowolne kombinacje technik, poprzez zalecenia, kiedy i jakie metody należy stosować, skończywszy na przepisach „krok po kroku"[7].

Dostawcy systemów ERP negocjują z klientami dwie wielkości: koszty i czas realizacji. Konsultanci szacujący oprogramowania posługują się takimi miarami jak roboczogodzina, roboczodzień czy roboczomiesiąc. Znając szacunek wielkości kosztów w jednostkach czasu pracy potrzebnego na realizacje, można przeliczyć go na wartość kosztu w walucie oraz na czas (terminy) realizacji, uwzględniając możliwość równoczesnej realizacji niektórych prac.

Niniejszy artykuł zawiera w rozdziale 1 opis fazy strategicznej procesu wdrożenia z uwzględnieniem jakości dostępnych informacji do wyceny. Kolejny rozdział jest przeglądem algorytmicznych metod szacowania. Rozdział 3 zawiera opisy metod niealgorytmicznych. Oba rozdziały zwracają uwagę na rodzaj i jakość danych potrzebnych do szacowania. W rozdziale ostatnim zaprezentowano wnioski wynikające z połączenia efektów poszczególnych etapów cyklu życia i danych potrzebnych do wycen. Na tej podstawie zaproponowano algorytm wykorzystujący ankietę do wyboru metody szacowania na poszczególnych etapach fazy strategicznej wdrożenia.

Znaczenie symboli umieszczonych na rysunkach 1,2,4 oraz w Dodatku B jest zgodne z notacją BPMN 2.0 [8].

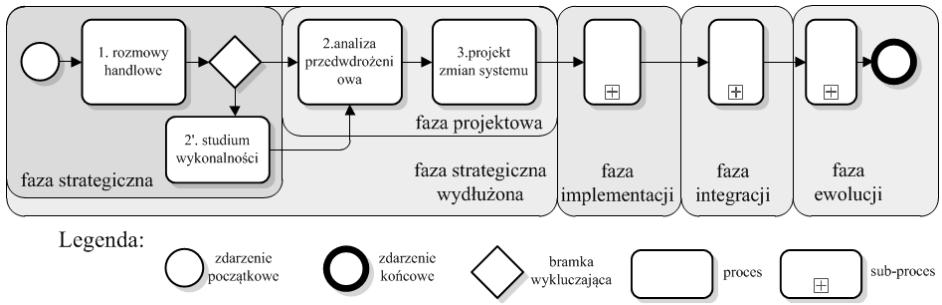
1. Etapy cyklu życia wdrażanego systemu ERP

Liczni autorzy opisują cykle życia oprogramowania, skupiając się na procesie produkcji oprogramowania lub tworzenia oprogramowania na zamówienie [3]. Wśród prezentowanych tam modeli żaden nie odpowiada w pełni procesowi wdrażania oprogramowania klasy ERP w średnim przedsiębiorstwie. Nie uwzględniają one „ruchomości" zakończenia fazy strategicznej (podpisania kontraktu) oraz możliwości wystąpienia dodatkowego etapu - studium wykonalności. Studium wykonalności nie jest istotne dla cyklu życia oprogramowania, ale dostarcza informacji do wyceny projektu.

Etapy fazy strategicznej są następujące:

1. wstępne rozmowy handlowe,
2. analiza przedwdrożeniowa
2'. studium wykonalności,
3. projekt zmian systemu,

Etapy fazy strategicznej oraz pozostałe fazy cyklu życia oprogramowania (implementacje, integrację, ewolucję) przedstawia Rysunek 1.



Rysunek 1. Etapy fazy strategicznej i pozostałe fazy wdrażania systemu ERP

Mając na uwadze szacowanie kosztów, należy pamiętać, że istotny w procesie sprzedaży jest moment podpisania kontraktu. Może to nastąpić po etapie 1., ale nie później niż po zakończeniu etapu 3. Okres ten nazywa się fazą strategiczną. W interesie dostawcy SI jest podpisanie umowy z klientem jak najszybciej, gdyż realizacja kolejnych etapów obciąża go kosztami, a ciągle istnieje ryzyko niepodpisania kontraktu. Jednakże wcześniejsze określenie kosztów w warunkach większej niepewności wiąże się z ryzykiem większego błędu szacowania.

1.1. Wstępne rozmowy handlowe

Dostawca odbywa spotkania z potencjalnym klientem w celu ustalenia zakresu i wartości kontraktu. Zwykle jest to jedno spotkanie wstępne i dwa lub trzy spotkania prezentacyjne. Niektóre z elementów zakresu prac zostają ujawnione szybko i szczegółowo. Dotyczą one przede wszystkim sprzętu komputerowego, wykonania infrastruktury sieciowej, licencji na poszczególne moduły systemu ERP. Niektóre elementy zakresu, np. modyfikacje SI wynikające ze specyficznych wymagań użytkowników, są trudne do określenia. Na tym etapie dostawca w niewielkim zakresie potrafi zidentyfikować wymagania klienta, których nie realizuje wersja standardowa sytemu ERP. Wiedza klienta o SI pochodzi na ogół z prezentacji handlowych, przez co nie zawsze potrafi on precyzyjnie określić te, które nie są standardowe. Wymagania, które dostawca jest w stanie pozyskać od klienta są zwykle niekompletne (brak jest wymagań, które klient uważał za nieistotne) i ogólne (klient nie jest w stanie ocenić istotnego

poziomu szczegółowości).

Jeśli dostawca potrafi wyspecyfikować ujawnione wymagania oraz zasugerować wymagania nieujawnione, może próbować wyceniać zmiany. Na przykład, klient określił wymaganie w obszarze produkcja, dotyczące zamawiania oddzielnie do każdego zlecenia produkcyjnego, surowców z grupy towarowej A. Wymaganie takie sugeruje nieujawnione wymagania dotyczące zmiany procesu tworzenia zamówień towarów w obszarze logistyka, gdzie z ogólnego planów zamówień dostaw wyłączyć należy zarządzanie surowcami z grupy A. Oba wymagania powinny być wycenione, mimo, ze ujawniono przez klienta tylko pierwsze z nich. Na tym etapie sporadycznie pojawiają się pojedyncze, szczegółowe wymagania typu: raporty różnie agregujące te same dane, wydruki w specyficznej formie używanej przez klienta.

1.2. Studium wykonalności lub analiza przedwdrożeniowa

Jeśli dostawca nie był w stanie wycenić dostosowania systemu (modyfikacje), musi przeprowadzić prace, dzięki którym ujawnione i uszczegółowione będą wymagania klienta. Realizuje wówczas studium wykonalności[9] lub analizę przedwdrożeniową. Mimo że oba rozwiązania mają doprowadzić do uszczegółowienia danych do wyceny, to cel utworzenia każdego z nich jest inny.

Studium wykonalności zawiera zestawienie zebranych informacji w postaci uporządkowanego dokumentu opartego na faktach gospodarczych[10]. Informacje dotyczą sfery ekonomicznej, organizacyjnej, technicznej. Celem studium jest określenie zakresu prac oraz kosztów realizacji przedsięwzięcia. Z dokumentu korzystają decydenci dostawcy, rozpatrując ekonomiczne aspekty realizacji projektu.

Analiza przedwdrożeniowa nie zawiera innych informacji niż te, które dotyczą systemu informatycznego w kontekście danego przedsiębiorstwa i realizacji prac. Wynikiem jest raport zawierający następujące elementy: zakres funkcjonalny wdrożenia, wykaz i opis procesów biznesowych, funkcji i danych zalecanych do uwzględnienia w zakresie funkcjonalnym wdrażanego systemu, zakres organizacyjny wdrożenia, proponowane cele wdrożenia, oczekiwane korzyści biznesowe, harmonogram prac [11]. Na tym etapie dostawca zakłada, że wymagania są kompletne i poziom szczegółowości odpowiada wymaganiom projektantów, dla których ten dokument jest podstawą dalszych prac.

Nawet w średnim przedsiębiorstwie produkcyjnym ewidencja wszystkich wymagań użytkowników i wszystkich procesów byłaby bardzo czasochłonna i kosztowna (od kilkuset do ponad tysiąca wymagań i procesów). Ponadto w większości pokrywałaby się z zapisami w dokumentacji systemu ERP. Dlatego też dostawcy wykonują analizę różnicową, która zawiera tylko te wymagania i opisy procesów, które nie są realizowane przez wersję standardową SI. Takie postępowanie znacznie skraca czas realizacji etapu, ale powoduje, że klient, żeby mieć obraz całego systemu, musi poznać dokumentację wersji standardowej łącznie z dokumentacją analizy przedwdrożeniowej.

Dostawca powinien zweryfikować jakość wymagań jakie zostały ujawnione na tym etapie pod kątem wykorzystania metod szacowania.

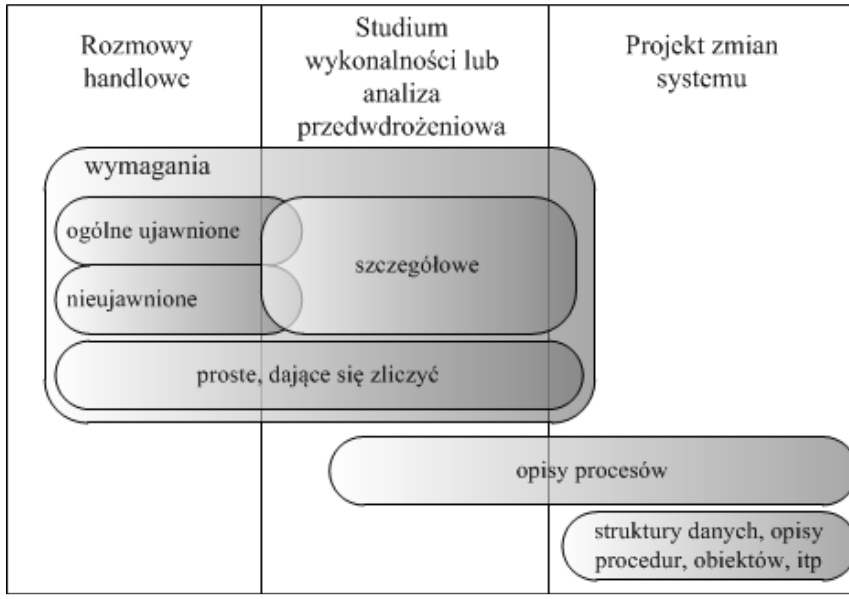
1.3. Projekt zmian systemu

Projekt systemu informatycznego to faza pośrednia między określeniem wymagań a realizacją. Dokumentacja, jaka powstaje, jest przeznaczona wyłącznie na użytek wewnętrzny dostawcy (działy programowania). W zależności od tego, jakimi metodami odbywać się będzie realizacja (programowanie strukturalne, obiektowe, zwinne, itp.), dokumenty projektowe zawierać mogą inne składowe [3]. Niektórzy producenci systemów ERP opracowali własne metodyki i w takim wypadku dokumentacja fazy projektowania będzie specyficzna. Przykładem może być metodyka Select Perspective [12] [13] lub ARIS [14]. Zawsze jednak występować będą wspólne elementy istotne dla procesów szacowania oprogramowania.

Pierwszym elementem prac projektowych jest uściślenie wymagań wynikających ze specyfiki realizacji. Poziom szczegółowości wymagań musi jednoznacznie determinować sposób ich realizacji. Poza tym dokumenty projektowe zawierają składowe opisujące struktury danych i procedury realizujące procesy. Istnieje wiele sposobów prezentacji informacji projektowej: od DFD, diagramy relacji encji, poprzez perspektywy modeli UML. Każdy z nich jest odpowiednim źródłem danych do szacowania realizacji oprogramowania.

1.4. Podsumowanie etapów cyklu życia

Wraz z kolejnymi etapami cyklu życia oprogramowania rośnie wiedza dostawcy o różnicach między procesami realizowanym w przedsiębiorstwie a funkcjonalnością oprogramowania standardowego. Na początku dysponuje on jedynie niekompletnym zbiorem wymagań ogólnych. W kolejnych etapach uzupełnia i uszczegóławia wymagania. Po etapie projektu dostawca może dodatkowo do szacowania użyć elementów projektowych takich jak: obiektów danych (tabel, pól), okien, interfejsów, itp. Z drugiej strony rosną koszty działań po stronie dostawcy. Jeśli dostawca podpisze kontrakt z klientem, koszty te znajdą się w wartości kontraktu, jeśli nie - obciążą w całości dostawcę. Informacje wejściowe, potrzebne do wykonania szacowania na poszczególnych etapach rozszerzonej fazy strategicznej przedstawia Rysunek 2.

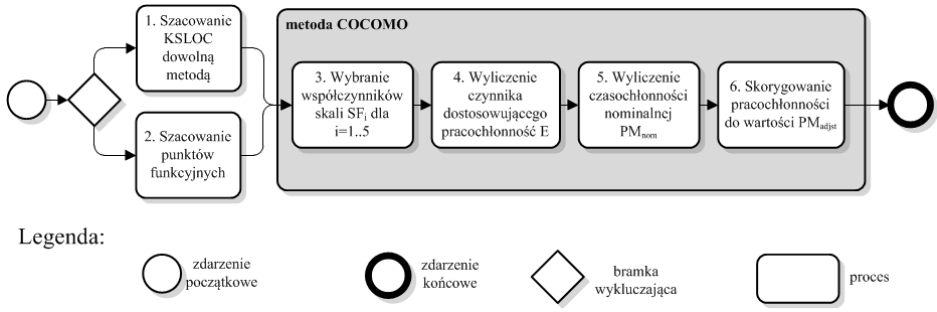


Rysunek 2. Informacje wejściowe dla wycen, na początkowych etapach cyklu życia

2. Algorytmiczne metody szacowania oprogramowania

2.1. Metoda COCOMO II

Metoda Constructive Cost Model (skr. COCOMO) została zaproponowana przez Barry Boehma w 1981 r. [15]. Od tego czasu powstało wiele wersji i odmian tej metody, np.: COCOMO81, COCOMO II [16]. Sekwencję procesów składających się na szacowanie przedstawia Rysunek 3. Za pomocą metody COCOMO można wyliczać czasochłonność (ang. Person per Month, skr. PM) na podstawie wielkości źródłowego kodu programu (ang. Kilo Source Line of Code, skr. KSLOC) (proces 1 na Rysunku 3). Informacje potrzebne do szacowanie ilości kodu pozyskane są z dokumentacji projektu SI. Ilości KSLOC przypisane są do elementów programu takich jak procedury, moduły, obiekty, itp. Ponieważ dla wielu współczesnych zastosowań ilość kodu często nie odpowiada czasochłonności, zmodyfikowano tę metodę, wykorzystując punkty funkcyjne (ang. Function Point, skr. FP) [17] (proces 2 na Rysunku 3) wyliczone na podstawie kompletnych i szczegółowych wymagań użytkownika.



Rysunek 3. Sekwencja procesów w metodzie COCOMO

Pierwszą czynnością jest określenie pięciu współczynników skali (ang. Scale Factor, skr. SF), których wartości wyznaczono empirycznie w pięciu klasach, w zależności od poziomu złożoności (od bardzo niskiej do bardzo wysokiej. Znając współczynniki skali, można wyznaczyć czynnik dostosowujący pracochłonność *E* (ang. Effort) zgodnie ze wzorem:

$$E = B + 0{,}01 \cdot \sum_{i=1}^{5} SF_i \tag{1}$$

gdzie: *B* jest stałą wynoszącą 0,91 dla modelu COCOMO II [17].

Wyliczenie pracochłonności nominalnej PM_{nom} odbywać się będzie zgodnie ze wzorem:

$$PM_{nom} = A \cdot (Size)^{E} \tag{2}$$

gdzie: *Size* - ilość linii kodu w jednostce KSLOC,

A - stała wyznaczona z historycznych projektów wynosząca 2,94 [17].

Dla modeli z pierwszych etapów cyklu życia oprogramowania (Application Composition Model, Early Design Model [17]) nominalny czas powinien zostać skorygowany siedmioma współczynnikami pracochłonności, zgodnie ze wzorem:

$$PM_{adis} = PM_{nom} \cdot \prod_{i=1}^{7} EM_i \tag{3}$$

gdzie: EM_i - (ang. Effort Multiplier) współczynniki pracochłonności.

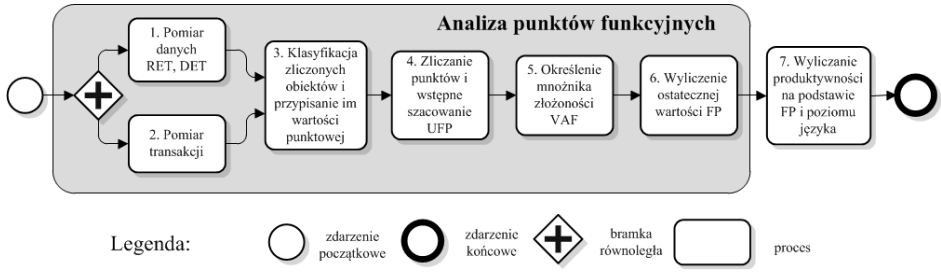
Dla modeli na kolejnym etapie cyklu życia (Post-Architecture Model), wzór na nominalną czasochłonność został rozszerzony o kolejnych 9 współczynników (i=1..16). Podobnie jak *SF* tak i *EM* wyznaczone zostały empirycznie. Dane potrzebne do wyliczeń SF i EM można znaleźć w dokumentacjach metody [17].

W literaturze możemy znaleźć wiele przykładów dostrajania metody *COCOMO* [18] [19] m.in. z wykorzystywaniem logiki rozmytej [20] [21] [22].

2.2. Szacowanie w oparciu o analizę punktów funkcyjnych

Metoda szacowania zaproponowana przez A.J. Albrechta [23] polega na wyliczeniu na podstawie szczegółowych wymagań, ilości punktów funkcyjnych (ang. Function Point, skr. FP) odpowiadających złożoności oprogramowania. Następnie za pomocą metody *COCOMO* lub *Szacowanie przez analogię* można przeliczyć ilość FP na czasochłonność lub koszty. Zbiór wymagań użytkownika będące podstawą obliczeń musi być kompletny a same wymagania szczegółowe.

Proces szacowania za pomocą punktów funkcyjnych przedstawia Rysunek 4.



Rysunek 4. Proces szacowania w oparciu o punkty funkcyjne

Metoda punktów funkcyjnych polega na wydzieleniu w wymaganiach lub gotowym programie pięciu klas obiektów (procesy 1 i 2 na Rysunku 4):

- logiczne wewnętrzne pliki (ang. Internal Logic File, skr. ILF),
- zewnętrzne interfejsy plików (ang. External Interface File, skr. EIF),
- zewnętrzne wejścia (ang. External Inputs, skr. EI),
- zewnętrzne wyjścia (ang. External Outputs, skr. EO),
- zewnętrzne zapytania (ang. External Inquires, skr. EQ.

Dwie pierwsze klasy opisują obiekty związane z danymi, trzy pozostałe - transakcje. Do szacowania na tym etapie (proces 3 na Rysunku 4) wykorzystuje się następujące wskaźniki:

- RET (ang. Record Element Type) - unikalna rozpoznawalna przez użytkownika podgrupa elementów danych w ILF lub EIF,
- DET (ang. Data Element Type) - unikalne, możliwe do zidentyfikowania przez użytkownika pole w ILF lub EIF,
- FTR (ang. File Type Referenced) - rozpoznawalny przez użytkownika zbiór logicznie powiązanych danych.

Następnie należy zidentyfikować wszystkie obiekty w klasach i przypisać im odpowiednią wartość wskaźników. ILF i EIF opisane są za pomocą RET i DET, natomiast EO, EI i EQ za pomocą FTR i DET. Na tej podstawie odczytujemy z tabeli wag [24] ilość nieostatecznych (nieskorygowanych) punktów funkcyjnych *UFP* (ang. Unadjustment Function Point) danego obiektu. Sumując wartości *UFP* dla wszystkich obiektów, we

wszystkich klasach, otrzymujemy wartość sumaryczną nieostatecznych punktów funkcyjnych.

Czynnik korygujący (ang. Value Adjustment Factor - *VAF*) uwzględnia wewnętrzną złożoność systemu niezwiązaną z jego funkcjonalnością. Wyznaczenie go polega na podaniu wartości wpływu dla 14 czynników, które mogą spowodować zwiększenie stopnia skomplikowania systemu (proces 5 na Rysunku 4). Listę czynników można znaleźć w dokumentacji metody [24]. Wartość *VAF* oblicza się wg wzoru:

$$VAF = B + 0{,}01\sum_{i=0}^{14} C_i \tag{4}$$

gdzie: B - empirycznie wyznaczona stała o wartości 0,65 [24],

C_i - wartość wpływu i-tego czynnika.

Znając VAF możemy obliczyć ostateczną wartość punktów funkcyjnych FP korygując wartość nieostatecznych punktów funkcyjnych UFP zgodnie ze wzorem:

$$FP = VAF \cdot UFP \tag{5}$$

Znając wartość FP, produktywność można wyznaczyć dwoma metodami:

- przeliczyć na KSLOC za pomocą empirycznie wyznaczonych wartości z tabeli przeliczeniowej [25] i dalej skorzystać z metody *COCOMO* do wyznaczenia czasochłonności,
- wykorzystując metodę *Szacowania przez analogię* może przeliczyć wartości FP bezpośrednio na czasochłonności, jeśli organizacja posiada odpowiednie dane historyczne,

Źródłem kompletnej i aktualnej dokumentacji *Analizy Punktów Funkcyjnych* jest strona WWW organizacji International Function Point Users Group [26].

3. Niealgorytmiczne metody szacowania oprogramowania

3.1. Dekompozycja i rekonstrukcja

Dekompozycja i rekonstrukcja to z powodu intuicyjności i uniwersalności bardzo popularna metoda. Stosowana jest w sytuacjach, kiedy szacowanie całego zakresu stanowi trudność, np. wynikającą z niejednorodności prac. W praktyce realizacji projektów informatycznych nieczęsto zdarzają się projekty, które można szacować z pominięciem tej metody.

Polega ona na podzieleniu szacowanego zakresu na wiele części. Sposób podziału może być dowolny i uzależniony od specyfiki projektu. Często dostawcy dokonują podziału za pomocą metody Work Breakdown Structure (skr. WBS) [13]. Po dokonaniu podziału części obiektu są szacowane lub dalej dzielone tą samą lub inną metodą. Głębokość podziału zależy od tego, jaką metodą odbędzie się szacowanie na kolejnym etapie. Mimo że w literaturze, metoda ta jest wymieniana na równi z innymi [4], to jej

rola w procesie szacowania jest różna od pozostałych. Szacowania projektów zaczyna się tą metodą, ale po dekompozycji wybierane są inne metody do szacowania cząstkowego. Szczegółowy opis metody dekompozycji zgodnie z WBS można znaleźć w wielu pozycjach literatury [27] [28] [29] [30].

3.2. Indywidualna ocena eksperta

Metoda szacowania poprzez *Indywidualną ocenę eksperta* to najczęściej stosowana metoda, nie tylko w praktyce tworzenia oprogramowania [31], ale i w innych przedsięwzięciach informatycznych, takich jak implementacje czy modyfikacje. Badania przeprowadzone w USA w 2002 roku pokazały, że 72% szacowań odbywa się tą metodą [32]. W pierwszym etapie polega na wytypowaniu ekspertów posiadających odpowiednią do zadań projektowych wiedzę i doświadczenie. W kolejnym etapie wyceny eksperci oceniają przydzielone im zakresy. W celu zmniejszenia błędów szacowania zmodyfikowano metodę o wykonanie szacowań parokrotnie dla różnych wariantów realizacji. Technika taka o nazwie PERT (ang. Program Evaluation and Review Technique) [27] [33] zakłada szacowanie dla: najbardziej korzystnego przypadku, najbardziej prawdopodobnego przypadku, najgorszego przypadku. Różni się jednak od znanej z analiz ścieżki krytycznej (m.in. CPM [34]) tym, że wykorzystywana jest jedynie do szacowania pojedynczych zadań. Po wcześniejszych procesach dekompozycji utracona została informacja o powiązaniach między zadaniami. Oczekiwana wartość szacowania wygląda wówczas następująco:

$$f(x) = \sum_{i=1}^{N} (Cp(x_i) + 4 \cdot Co(x_i) + Ck(x_i))/6 \tag{6}$$

gdzie: Cp – najbardziej korzystna wartość i-tego zadania,
Co – najbardziej prawdopodobna wartość i-tego zadania,
Ck – najmniej korzystna wartość i-tego zadania.

Dokładność wyników zależy wyłącznie od doświadczenia eksperta. Kryteria wyboru eksperta są nieprecyzyjnie określone. Wpływ osobowości jest na tyle duży, że większe doświadczenie nie gwarantuje dokładniejszych wycen. Zdarzają się eksperci zwykle zawyżający szacowania, zaniżający lub nieprzewidywalni.

Metoda ta możliwa jest do stosowania od etapu pierwszych kontaktów dostawcy z klientem. Dzięki odpowiedniemu doborowi ekspertów szacować można nawet na podstawie niekompletnego zbioru ogólnych wymagań użytkownika.

3.3. Ocena eksperta w grupach

Metoda polega na przedstawieniu do wyceny tego samego zakresu prac więcej niż jednemu ekspertowi. W wersji niestrukturalnej (recenzje grupowe) eksperci wspólnie ustalają wartość wyceny lub zakres wyceny. W wersji strukturalnej zwanej Wideband Delphi [35] [15] ustalenia ekspertów dokonuje się w sposób sformalizowany i efektem jest

wycena punktowa.

Praca w grupach jest kosztowniejszą metodą niż praca pojedynczych ekspertów, ale przewagą nad *Indywidualną oceną eksperta* jest zmniejszenie wpływu składnika osobowościowego. Mimo różnych doświadczeń, charakterów i skłonności, albo eksperci dojdą do wspólnych ustaleń, albo jak w odmianie Wideband Delphi rozstrzygnięcie spornych wycen nastąpi poprzez przypisanie umownych punktów.

Metoda szacowania stosowana jest szczególnie często w początkowych etapach projektów informatycznych, w sytuacjach dużej niepewności i nieprecyzyjności wymagań.

3.4. Zliczanie, obliczanie i ocenianie

Metoda polega na odszukaniu w projekcie obiektów, które dają się zliczyć, takich jak wymagania, funkcje, przypadki użycia, historyjki, raporty, okna dialogowe, tabele baz danych, klas. Do każdego zidentyfikowanego obiektu wymagane jest przypisanie wielkości składowej szacowania (kosztu lub czasu). Wartości szacowane są funkcją obiektów składających się na projekt informatyczny:

$$f(x) = \sum_{i=1}^{N} C(x_i) \tag{7}$$

gdzie: x - zliczony obiekt,

N - ilość zliczonych obiektów,

C - oceniany koszt zliczonego obiektu.

Metoda może być stosowana na każdym z etapów powstawania lub modyfikacji oprogramowania. Nie jest skomplikowana pod warunkiem, że w dokumentacji źródłowej, np. w studium wykonalności lub analizie przedwdrożeniowej, można wyodrębnić zliczane obiekty. Wadą jest duże ryzyko pominięcia obiektów lub zakresów prac, które mają wpływ na wartość całego projektu, na przykład pominięcie tabel pomocniczych lub nieuwzględnienie kosztów przygotowania zapytań filtrujących w trakcie szacowania kosztów okien interfejsów. Ważnym etapem tej metody jest ocena poszczególnych obiektów. Dokonać tego można, wspomagając się metodą *Indywidualnej oceny eksperta* lub *Oceny eksperta w grupach*. Metoda jest efektywna w projektach, w których udało się wyodrębnić niewiele rodzajów obiektów za to występujących licznie, na przykład 30 raportów, 25 zapytań SQL i 18 okien interfejsów.

3.5. Szacowanie przez analogię

Metoda polega na podzieleniu projektu na takie części, jakie występują w już zrealizowanym projekcie. Szacując wybrane części, można obliczyć stosunek wielkości obu projektów (nowego i zrealizowanego). Znając relacje między wielkościami i koszty zrealizowanego projektu, można oszacować wartość nowego projektu.

Trudność polega na zebraniu danych historycznych z projektów o podobnym charakterze i strukturze jak szacowny projekt. Dodatkowym problemem jest wybór reprezentatywnej części zdekomponowanego projektu, na podstawie której obliczany jest współczynnik krotności. Pominięcie istotnych obiektów może zwiększyć błąd szacowania.

Danymi wejściowymi dla tej metody powinny być obiekty danych i programów (okna, zapytania SQL, FP). Wykorzystanie wymagań, nawet szczegółowych, nie pozwoli na obliczenie współczynnika krotności, a co za tym idzie całego szacowania. Stąd metoda może być stosowana dopiero na etapie, na którym znane są efekty fazy projektowania.

3.6. Szacowania oparte na zastępstwie

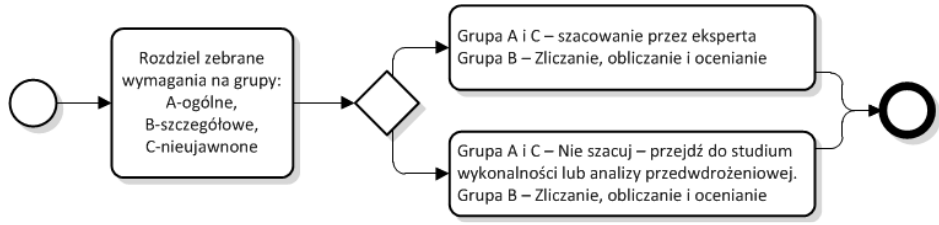
Metoda ta, podobnie jak poprzednia, wymaga znajomości kosztów wcześniej zrealizowanych w danej organizacji obiektów standardowych (interfejsy, raport, itp.). W zależności od wersji metody obiekty mogą być różnie grupowane. Na przykład, Putnam [33] i Humphrey [36] wyodrębnili klasy obiektów: bardzo małe, małe, średnie, duże, bardzo duże. Innym sposobem klasyfikacji jest metoda standardowych składników [4] używana do szacowania oprogramowania obiektowego. Jeśli dostawca SI wykorzystuje programowanie ekstremalne lub bliskie metodom Agile [37], standardowym składnikiem mogą być tzw. historyjki.

Następnie grupom obiektów przypisuje się średnie historyczne miary kosztów, np. liczby linii kodu (skr. ang. LOC), roboczogodziny lub roboczodni. W podobny sposób trzeba sklasyfikować obiekty z nowego projektu. Wówczas, na tej podstawie można obliczyć ich sumaryczną wartość.

Podobnie jak poprzednia metoda, ta również powinna być wykorzystywana. gdy znane są klasy obiektów programistycznych. Wyjątkiem są organizacje używające programowania ekstremalnego, czy zwinnego. W tym przypadku, koszty „historyjek" jakie udokumentowane zostały na etapie rozmów z klientem mogą być zastępowane danymi historycznymi. Praktyka wycen [2] wskazuje jednak, że wykorzystywana może być na etapie wcześniejszym (analiza przedwdrożeniowa), kiedy znane są tylko wymagania szczegółowe.

4. Wnioski

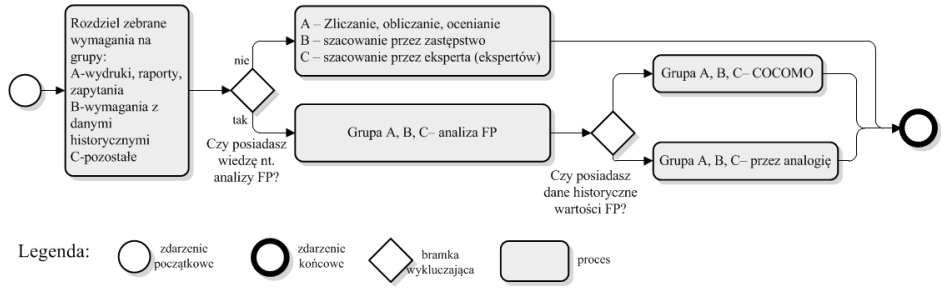
Podsumowując, należy zauważyć, że z jednej strony realizacja pierwszych etapów cyklu życia oprogramowania dostarcza coraz więcej informacji o planowanym wdrożeniu, a z drugiej strony istnieje zbiór metod szacowania, które na wejściu potrzebują różnego typu informacji (w zależności od metody).



Rysunek 5. Wybór metod szacowania na etapie rozmów handlowych

Na podstawie dokonanej analizy metod szacowania, autorzy proponują własną metodę wyboru najtrafniejszego sposobu oceny kosztów i czasu realizacji wdrożenia (modyfikacji oprogramowania w trakcie wdrożenia).

Dla etapu rozmów handlowych algorytm postępowania przedstawiony jest na Rysunku 5. Jak widać dla wszystkich grup wymagań szacować można czas i koszty tylko w przypadku gdy wymagania te będą już kompletne. W praktyce należy dla większości wymagań, przejść do studium wykonalności (analizy przedwdrożeniowej). W taki przypadku mamy do wyboru inne metody szacowania co przedstawiono na Rysunku 6.



Rysunek 6. Wybór metod szacowania na etapie analizy przedwdrożeniowej (studium wykonalności)

Etap projektu zmian systemu dostarcza wyceniającym, oprócz wymagań dodatkowych, informacji związanych z realizacją prac - struktury danych, informacje o procedurach realizujących procesy, obiektach, itp. Podobnie, jak na poprzednich etapach, dostawca powinien dokonać klasyfikacji dostępnych danych:

A. dane umożliwiające szacowanie obiektów w KSLOC,
B. dane, o których lub podobnych dostawca posiada informacje historyczne (koszty realizacji),
C. dane, o których dostawca posiada sklasyfikowane informacje historyczne (koszty realizacji),

Następnie wymagania typu A dostawca szacuje metodą *COCOMO* po wcześniejszym oszacowaniu KSLOC, typu B - *Metodą przez Analogię*, typu C - *Metodą przez Zastępstwo*.

Powyższe propozycje postępowania pozwalają wykorzystywać te metody, które w warunkach poszczególnych etapów będą najefektywniejsze.

Bibliografia

1. M. Burns, „How to select and implement an ERP System," 2005. [Online]. Available: http://www.180systems.com/ERPWhitePaper.pdf.
2. P. Plecka, „Selected Methods of Cost Estimation of ERP Systems' Modyfications.," *Zarządzanie Przedsiebiorstwem,* p. w druku, 2013.
3. I. Sommerville, Software Engineering, Edingurgh: Pearsom Education Limited, 2007.
4. S. McConell, Software Estimation: Demystifying the Blac Art., Microsoft Press, 2006.
5. R. Meli, „Early Function Points: a new estimation method for software projekct," w *WSCOM97*, Berlin, 1997.
6. L. Santillo, M. Conte i R. Meli, „Early &Quick Function Point: Sizing More with Less," w *Metrics 2005, 11 th IEEE Intl Software Metrics Symposium*, Como, Italy, 2005.
7. B. Boehm, C. Abts i S. Chulani, „Software Development Cost Estimation Approaches – A Survey," *Annals of Software Engineering,* tom 10, nr 1-4, pp. 177-205, 2000.
8. „BPMN," Object Management Group, 2013. [Online]. Available: http://www.bpmn.org/.
9. K. Frączkowski, Zarządzanie projektem informatycznym., Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2003.
10. J. Philips, IT Project Management. On Track from Start to Finish., Osborne, 2004.
11. K. Justynowicz, „Analiza przedwdrożeniowa coraz popularniejsza," wrzesień 2007. [Online]. Available: http://www.bcc.com.pl/akademia-lepszego-biznesu/analiza-przedwdrozeniowa-coraz-popularniejsza.html.
12. P. Allen i S. Frost, Component-Based Development for Enterprise Systems, Applying the Select Perspective, Cambridge: Cambridge University Press, 1998.
13. „Select Business Solution," [Online]. Available: http://www.selectbs.com.
14. „ARIS," [Online]. Available: http://www.softwareag.com.
15. B. Boehm, Software Engeneering Ecomonics, New York: Englewood Clifs, 1981.
16. B. W. Boehm, Software Cost Estimation with COCOMO II, Prentice Hall, 2000.
17. J. Baik, COCOMO II, Model Definition Manual, Version 2.1, Center for Software Engineering at the University of Southern California, 2000.
18. J. Hele, A. Parrish, B. Dixon i R. Snith, „Enhancing the Cocomo estimation models," *Software, IEEE, Volume:17, Issue: 6* , pp. 45-49, 2000.

19. S. Aljahdali i A. Sheta, „Software effort estimation by tuning COOCMO model parameters using differential evolution," w *Computer Systems and Applications (AICCSA), IEEE/ACS International Conference on*, Hammamet, 2010.
20. Z. Fei, „f-COCOMO: fuzzy constructive cost model in software engineering," w *Fuzzy Systems, 1992., IEEE International Conference on*, San Diego, CA, 1992.
21. R. C. Satyananda, „An Improved Fuzzy Approach for COCOMO's Effort Estimation using Gaussian Membership Function," *JOURNAL OF SOFTWARE,* tom VOL. 4, nr NO. 5, pp. 452-459, July 2009.
22. I. Attarzadeh, „Improving estimation accuracy of the COCOMO II using an adaptive fuzzy logic model," w *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, Taipei, 2011.
23. A. Albreht, „Measuring Application Development Productivity," w *Proceedings of the Joint SHARE, GUIDE, and IBM Application Development Symposium*, Monterey, California, USA, 1997.
24. IFPUG, Function Point Counting Practices: Manual Release 4.1, Westerville, OH: IFPUG, 1999.
25. „The QSM Function Points Languages Table," QSM, 2013. [Online]. Available: http://www.qsm.com/resources/function-point-languages-table.
26. „International Function Point Users Group," [Online]. Available: http://www.ifpug.org/.
27. R. D. Stutzke, Estimation Software-Intensive Systems, Upper Saddle River, New York: Addison-Wesley, 2005.
28. R. Tausworthe, „The work breakdown structure in software project management," *Journal of Systems and Software,* tom 1, 1984.
29. E. Norman, S. Brotherton i R. Fried, Work Breakdown Structures: The Foundation for Project Management Excellence, John Wiley & Sons, 2010.
30. G. Haugan, Effective Work Breakdown Structures, Project Management Institute, 2002.
31. M. Jorgensen, „A Review of Studies on Expert Estimation of Software Development Effort," *Journal of Systems and Software,* tom 70, nr 1-2, p. 37–60, Fabruary 2004.
32. B. Kitchenham, S. L. Pfleeger, . B. McColl i S. Eagan, „An empirical study of maintenance and development estimation accuracy.," *Journal of Systems and Software,* tom 64, nr 1, pp. 57-77, 2002.
33. P. L. H. i. W. Myers, Measures for Excellence: Reliable Software on Time, Within Budget, Englewood Cliffs, NY: Yourdon Press, 1992.
34. J. W. Fondahl, „The History of Modern Project Management Precedence Diagramming Methods: Origins and Early Development," *Project Management Journal,* tom XVIII., nr 2, 1987.
35. NASA, „ISD Wideband Delphi Estimation," 09 2004. [Online]. Available:

http://software.gsfc.nasa.gov/assetsapproved/PA1.2.1.2.pdf.
36. W. S. Humphrey, A Discipline for Software Engineering, Addison Wesley, 1995.
37. M. Cohn, Agile Estimating and Planning, Upper Side River, NY: Prentice Hall PTR, 2005.

Streszczenie

W pracy poruszono problem doboru metod szacowania kosztów modyfikacji systemu ERP podczas wdrożenia. Przeprowadzono przegląd dostępnych metod opisanych w literaturze oraz scharakteryzowano etapy fazy strategicznej procesu wdrożenia. Na podstawie analizy zakresu danych wymaganych przez poszczególne metody oraz danych uzyskiwanych na różnych etapach, zaproponowano algorytmy doboru metod do tychże etapów.

Słowa kluczowe: ERP, projekt informatyczny, szacowanie kosztów, analiza punktów funkcyjnych, wdrożenia systemów informatycznych.

Summary

The work discusses the problem of selecting methods for valuing the costs of modifying ERP systems during implementation process. The methods presented in literature have been reviewed and the stages of strategic phase of implementation have been characterised. On the basis of the analysis of data required by each method and the data obtained at different stages, algorithms of method selection for each stage ware proposed.

Key words: ERP, IT project, cost valuation, function points analysis, IT systems implementations.