Anna BARTKOWIAK[1 2], Radoslaw ZIMROZ[3]

University of Wroclaw (1), Wroclaw School of Applied Informatics (2), Wroclaw University of Technology (3)

Curvilinear dimensionality reduction of data for gearbox condition monitoring

Abstract. *Our aim is to explore the CCA (Curvilinear Component Analysis) as applied to condition monitoring of gearboxes installed in bucket wheel excavators working in field condition, with the general goal to elaborate a probabilistic model describing the condition of the machine gearbox. To do it we need (a) information on the shape (probability distribution) of the analyzed data, and (b) some reduction of dimensionality of the data (if possible). We compare (for real set of data gathered in field conditions) the 2D representations yielded by the CCA and PCA methods and state that they are different. Our main result is: The analyzed data set describing the machine in a good state is composed of two different subsets of different dimensionality thus can not be modelled by one common Gaussian distribution. This is a novel statement in the domain of gearbox data analysis.*

Streszczenie. *W pracy przedstawiono wyniki prac nad zastosowaniem CCA (Curvilinear Component Analysis - analiza komponentów krzywoliniowych) do nieliniowej redukcji wymiarowości danych wykorzystywanych do diagnostyki przekładni planetarnej stosowanej w układach napędowych koparki kołowej. Do oceny stanu technicznego niezbędne jest zbudowanie modelu probabilistycznego zbioru cech diagnostycznych. Modelowanie danych wielowymiarowych (gęstości prawdopodobieństwa) dla wszystkich wymiarów jest trudne, i ze względu na istniejącą redundancję, nieuzasadnione, dlatego prowadzi się badania nad redukcją wymiarowości zbiorów cech diagnostycznych. W artykule porównujemy dwuwymiarowe reprezentacje zbioru cech uzyskane metodami CCA i PCA (analiza składowych głównych) wykazując różnice w uzyskanych wynikach. Głównym wynikiem pracy jest identyfikacja w przestrzeni cech diagnostycznych dla przekładni w stanie prawidłowym dwóch podzbiorów danych o różnej rzeczywistej wymiarowości zatem nie mogą być one modelowane za pomocą jednego modelu o charakterystyce gaussowskiej. Interpretacja tych podzbiorów wiąże się z występowaniem różnych obciążeń maszyny.* (***Redukcja wymiarowości danych przy monitorowania stanu skrzyni biegów***)

Keywords: Vibration signals, Power spectra, Dimension reduction, Non-linear mapping, Self-organizing network
Słowa kluczowe: sygnały drganiowe, widma sygnałów, redukcja wymiarowości, mapowanie nieliniowe, sieci samoorganizujące się

Introduction

Our aim is to investigate the curvilinear component analysis used in condition monitoring of gearboxes when considering vibrations emitted by these gearboxes. The problem is very important and complicated, because the state of a gearbox can not be evaluated directly basing on raw vibration time series and one should use more advanced methods providing a multidimensional description of gearbox condition. One such method is spectral representation of vibration signals in the form of power spectra. This takes us on the ground of multivariate data analysis. The analysis of vibration signals based on their spectral decomposition is well known [5, 7, 9, 12]. However, the applied methods are mostly those working on covariances or correlations between the observed features, which implies using linear models. There is some evidence that the dependencies among the observed variables may be curvilinear, which is not captured by ordinary, Pearsons' correlations, thus non-linear methods should be used to capture the true dependencies. One such method is CCA, Curvilinear Component Analysis, developed by Demartines and Herault [3, 6]. The method works on inter-point distances using specific cost function giving favour to inter-point proximities in the output space. When working locally (that is, with small neighbourhoods), CCA permits to unfold the non-linear structures of the data yielding as output some flat manifolds of lower dimension. It allows also to make projections of the original data to the obtained manifold of lower dimension 'dim'. In case of dimension 'dim' equal to 2 or 3 it is possible to visualize the data in 2D or 3D scatterplots. Our work presented here may be viewed as a case study of applying CCA to vibrations data recorded from gearboxes, containing amplitudes of 15 power spectra, which constitute 15 derived features. Our specific goal is to investigate the shape of the bivariate projections of amplitudes of the power spectra derived by using the Matlab PSD (power spectral density) function. This is done with the intention to obtain a 2D representation of the data in a plane and build there a decision boundary delimiting the 'normal state' of a device from an 'abnormal' one [2]. In the paper we analyze data obtained by [1] for a machine being in a good state; the respective data set is called set B (good). We state that the distribution of the data is decidedly not Gaussian. We compare also the projections by PCA and CCA and discuss their similarities and specificities. This is our original contribution. Our intention is to look more closely at the 2D projections and determine their shape (distribution) for devices working under small/no work load and under typical load. This is important for building models of the data. In next section (2) we shortly introduce the data used for analysis. Section 3 discuses the results (shape of the distribution, relevance of the 2D projection, fraction of explained total variance) obtained for the analyzing data when using the PCA method. Section 4 introduces the CCA method and shows the 2D projections obtained by this method. Simultaneously analogous projections from PCA are shown. The possibility of estimating the intrinsic dimensionality of the data is investigated. In Section 5 some summary of the results and conclusions are presented

Data used for the analysis

We use part of the data gathered and analyzed by Bartelmus and Zimroz [1]. The data are given in the form of rectangular matrices of size $n \times d$, with n denoting the number of rows and d the number of columns of the data matrix $\mathbf{X}$. A brief description of the data might be as follows (for details, see [1]). Vibration signal were recorded for a planetary gearbox (called in the following also device B or machine B) being in good condition (device B). The vibration signals were recorded during ca. 15 minutes for device B, which resulted in n_B = 951 segments expressing vibration signals gathered in time moments when the respective devices were working in time varying load conditions. Each of the segments was subjected to the Discrete Fourier Transformation (function PSD from Matlab), which yielded finally 15 power spectrum components (considering only real part of the spectrum). In such a way 15 variables named *pp's* were obtained; records of these variables constitute data matrices $\mathbf{A}$ and $\mathbf{B}$, the last being the subject of our analysis. Thus each data vector (instance, segment) contains values of d = 15 variables,

named $pp_1, \ldots, pp_{15}$; it may be viewed at the same time as a d-dimensional data point located in the d-dimensional Euclidean space.

Linear reduction of dimensionality using PCA

PCA (Principal Component Analysis) is one of the most frequently used methods for reduction of multivariate data and its denoising [5, 7, 9-12]. Using eigenvectors of the covariance matrix (or correlation matrix) of a data matrix $\mathbf{X}$ of size $n \times d$, one projects the data vectors (called also data points) to a lower dimensional subspace of dimension say k. This yields k new variables, called principal components and denoted as $PC_1, \ldots, PC_k$. The derived PCs have a number of favourable properties [5, 7]. How to find the proper dimension k? In [9] it was found that taking the correlation matrices for the performed analysis, the proper dimension for the set B is k=2 or k=3. Now we have investigated how the original variables $X_1, \ldots, X_{15}$ are reproduced when taking k=2, k=3, k=6 and k=10 PC variables. Before performing the calculations, the data matrix $\mathbf{X}$ was standardized to have column means equal to 0 and standard deviation (of each column) equal to 1. The PCA was performed using correlation matrix of the standardized data. We found, for example, that the original variables $X1$ and $X2$ are reproduced for dimensions k = 2 and k = 3 quite satisfactorily (reproduced fractions: for k = 2: 0.88 and 0.82, for k = 3: 0.88 and 0.84 appropriately), while the original variable $X3$ is reproduced quite poorly (reproduced fractions for k = 2 and k = 3 are 0.33 and 0.54 appropriately). The constructed features PC_1 , PC_2 and PC_3 permit to make a 2D and 3D visualization of the data. Figure 1, right exhibit, shows 2D projections of the analyzed data for four samples of size n = 300 each. One may notice that the shape of the projected data is quite peculiar: All the four projections show clearly that the analyzed data set B is composed from 2 types of points: those corresponding to time instances when the excavator has worked under small or no load (points marked in black), this state will be called in the following NO LOAD); and those working under normal load condition (points marked in red). These two types of points constitute two subgroups of the data referred up from now as NO LOAD and LOADED type. The main result of this section is that the data, viewed as projections obtained by PCA, constitute a mixture of two different types of data points. This indicates that the probability density of the data is not a common normal (Gaussian) distribution.

CCA – Curvilinear Component Analysis

General concept – obtaining projections to lower dimensionality manifolds

The curvilinear analysis (CCA) was introduced by Demartines and Herault [3, 4], see also [6, 8] for further applications of the method. The basic assumption underlying this method is that the multivariate data with d variables are located truly in a manifold of lower dimension, say p ($p < d$), moreover this subspace is somehow folded, which makes the relations between the observed variables – when viewed in R^p – to appear as non-linear ones. The problem to solve is formulated as follows:

We have N data vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$, each vector is located in d-dimensional data space R^d , that is $\mathbf{x}_i \in R^d$, $i=1, \ldots, N$. This space is called the input space. Each data point has d components constituting observed values of the variables $X_1, \ldots, X_d$. The main idea is to find a mapping of the given N points to a lower dimension subspace R^k ($k < d$) called the output space. The obtained projections in R^k will be denoted as $\mathbf{y}_i$, $i = 1, \ldots, N$,.

How to find a proper mapping? This can be done in many ways. One possibility is the following one: For every pair of points (i, j), $i \neq j$, take the inter-point distances X_{ij} in the input space and – basing on some criterion E expressing 'error' or 'cost'– find the corresponding inter-point distance Y_{ij} in the output space. Again, the distance between two points (i, j) may be defined in many ways, the simplest and most popular is the Euclidean distance. To find the proper mapping one needs to solve an optimization problem: namely to find values of the $\mathbf{y}_i$-s that minimize the assumed error function E.

Also the error function E can be defined in many ways. For example, the Sammon stress function is well known in this context, it is defined as

$$E_s = \frac{1}{\sum_{i,j} X_{ij}} \sum_{i,j} \frac{(X_{ij} - Y_{ij})^2}{X_{ij}} \tag{1}$$

The computational algorithm elaborated by Sammon (see reference in [4], p. 141) has the following drawbacks: a) the computational complexity grows exponentially with N; b) the solution is valid only for given N points $\mathbf{x}_i$; to make the mapping for a new point $\mathbf{x}_0$, the new point should be added to the set containing the N points, and all calculations should be performed anew with the set of the N + 1 data points; c) it aims at to preserve the distances in the input space and fails (yields bad mapping) when the data cloud is strongly U-shaped.

Demartines and Herault [3] tackle the problem of mapping in a different way.

1. Their input space contains N data points $\mathbf{x}_i$ which constitute de facto representatives of the entire data. The representatives are obtained from a process of VQ (Vector Quantization) of the entire data space; if the data set is of small size, then all data points may be taken as representatives.
2. The error (or cost) function is defined as

$$E = \frac{1}{2} \sum_i \sum_{j \neq i} (X_{ij} - Y_{ij})^2 F(Y_{ij}, \lambda_y) \tag{2}$$

   The function $F(Y_{ij},\lambda_y)$ is chosen as a bounded and monotonically decreasing function, in order to favor local topology conservation (as in SOM [11]).
3. The mentioned authors [3, 4] elaborated a fast iterative algorithm yielding the mappings of $\mathbf{x}_i$ to $\mathbf{y}_i$, $i = 1, \ldots, N$. For fixed i, the updating of the value $\mathbf{y}_j^{(r)}$ from the r^{th} iteration is done as $\mathbf{y}_j^{(r+1)} = \mathbf{y}_j^{(r)} + \Delta \mathbf{y}_j$, where $\Delta \mathbf{y}_j$ is given by the formula (the coefficient $\alpha(r)$ denotes the learning coefficient at iteration no. r)

$$\Delta \mathbf{y}_j = \alpha(r) F(Y_{ij}, \lambda_y)(X_{ij} - Y_{ij}) \frac{\mathbf{y}_j - \mathbf{y}_i}{Y_{ij}}, \forall j \neq i \tag{3}$$

We have applied the CCA method to the data set B described in section 2. For calculations we have used the Matlab function cca from the Matlab SOM Toolbox (www.cis.hut.fi/projects/) using default settings of the function. The elaborated set contains 951 data vectors, with 15 components each. Our calculations were performed on samples counting n=300 elements. Each sample, before entering the CCA procedure, was standardized to have column means equal 0 and column variances equal 1. We have made projections to k = 2 dimensional subspaces. Next we have visualized the obtained projection in a scatter plot. For each of the samples we have constructed two scatter plots: one exhibiting the projections obtained by CCA, and the other exhibiting analogous projections using the first two principal components. The obtained scatter plots are shown in Figure 1. The displayed results were

obtained in the following framework: sample size: n = 300, starting from: PCA, number of epochs: 10.

We have made a lot of other simulations investigating the effect of sample size and number of epochs (iterations) and starting values used by the algorithm CCA. Because of lack of space, we do not show them here.

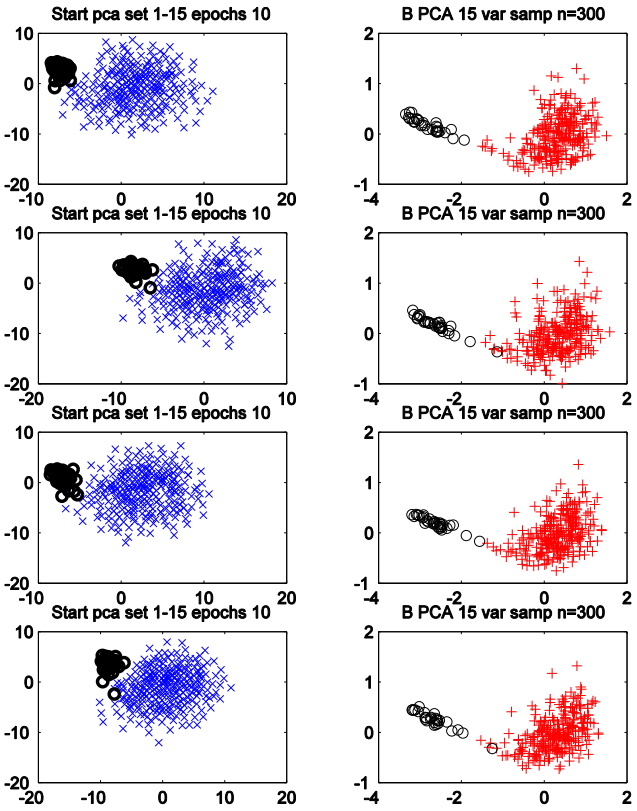

Fig.1. Projections of four samples counting n=300 data points each. Left: CCA with start from PCA. Right: classical PCA.

General remarks on the work of the CCA method. CCA starts from the supposition that the data are located in a lower dimension manifold which is somehow folded, and due to some additional noise appears to be located in the global d-dimensional space R^d of observed variables. There are two stages of the algorithm: first, we need a global unfolding of the average manifold, and second, we need a local projection of the data onto their average manifold. During the unfolding only local information on the data distances is needed. This means that the algorithm - at the given stage - takes into considerations only data pairs (i, j), with distances Y_{ij} relatively near each other. This is realized by defining appropriately the function $F(Y_{ij}, \lambda_y)=F_\lambda(Y_{ij})$ appearing in the cost function (2). It is usually defined as:

(4) $$F_\lambda(.) = 1 \text{ for } Y_{ij} < \lambda \text{ ; and =0 for } Y_{ij} \geq \lambda$$

This has the effect that (quoting [6]) only some of the error terms in formula (2) need to be minimized: those for which the distance Y_{ij} is smaller than some predefined λ. Thus, allowing the matching for only short distances is a way to respect the local topology. It has been proved [4] that this condition, applied to the output distance, ensures a global unfolding much better than other mapping techniques, which apply it to the input distances. See [6] for details of the second phase. Generally, also from other simulations not shown here, we stated that results do not depend from the size of the samples and number of epochs used for the simulations, but they depend crucially from the starting values of the $\mathbf{y}_j$ - s used by the cca algorithm. Let us mention that the starting values should be obtained from a kind of VC (vector quantization) of the space R^d, which was advised, for example, in [6]. Our experiments show clearly that start from PC values has a good effect, while starting from completely random initial values is bad (has difficulties with global convergence) and should be avoided.

Finding intrinsic dimension of the data

To see the effect of the unfolding and projection to a lower-dimensional manifold, Demartines and Herault [3, 6] proposed to use the so called dydx plots (distance of y-s and distance of x-s). These are simply scatter plots exhibiting for each pair (i,j) of data points the distances $(Y_{ij}; X_{ij})$ taking Y_{ij} as the 'x' coordinate and X_{ij} as the 'y' coordinate. Each scatter plot contains also the diagonal line $y = x$.

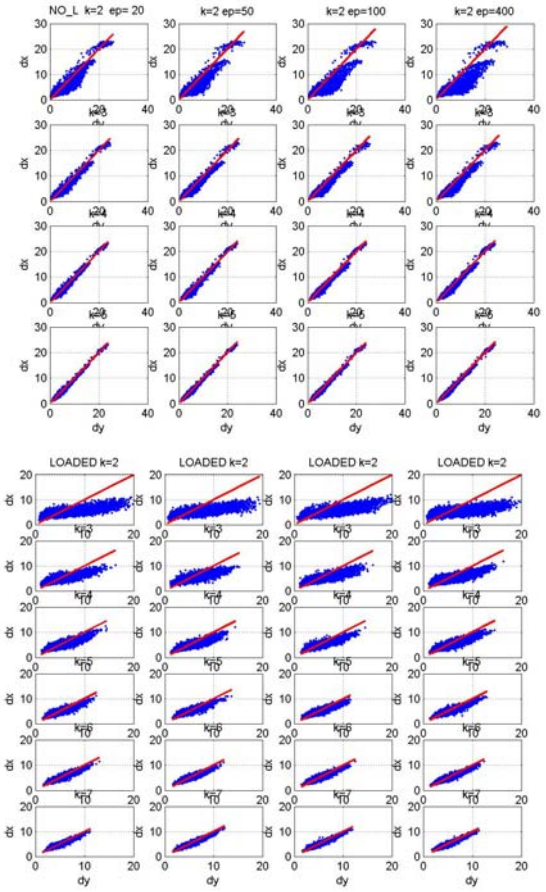

Fig.2: Seeking for intrinsic dimension by inspecting dydx plots constructed for the subsets called NO_LOAD (top) and LOADED (bottom). Notice the difference in the displays for the two subgroups. All the dydx plots are based on samples of the same size.

When seeking for the intrinsic dimensionality of the input data, we may apply the following principle: if the distribution of the data points lies on the diagonal, we can lower the output dimension. On the other hand, when the distribution becomes thick, the output dimension is too small. In case of multi modal input it is interesting to construct the dydx plots separately ([6], p. 632).

Our data proved to be composed of two different subgroups (see Figure 1). The first subgroup (with n = 104), consists of data points characterizing the NO LOAD status of the machine. The second subgroup, a larger one, (with n = 847), consists of data points characterizing the LOADED status of the machine.

The upper exhibit in Figure 2 corresponds to the NO LOAD status of the investigated machine. We investigated the dimensionality equal k = 2; 3; 4; 5; 6. For each k we have run cca with PCA as starting point and four different numbers of epochs: 20, 50, 100 and 400. The appropriate dydx plots are shown in the top exhibit of Figure 2. Rows of that exhibit correspond to different values of k (that is, dimensionality), and columns to different number of epochs.

What concerns the LOADED status, the investigations were done also for k = 2; 3; 4; 5; 6; 7. To make a balance for the upper display, for each k, four samples of size n = 104 were sampled from the LOADED subgroup. For each combination k *x sample,* we have run the function cca using PCA as starting point. In such a way 6 x 4 dydx plots were obtained - they are shown as subplots in the bottom exhibit in Figure 2. Rows in that exhibit correspond to different values of *k,* and columns to different samples.

Comparing the two exhibits in Figure 2 one states enormous difference in the estimated dimensionality of the samples (note, all samples in that figure are of the same size n = 104). The intrinsic dimension of the LOADED group seems to be at least equal 4 and that of the NO LOAD group only 2.

The main result of this section is: The subgroups NO LOAD and LOADED are of different dimensionality.

Discussion of the results and closing remarks

We have investigated the algorithm CCA (Curvilinear component analysis) [3] as applied to set B containing data from machine being in good condition. This is the first time in the context of gearbox condition monitoring, when the CCA method was applied. Summary of the results:

1. The algorithm is an iterative one. For the analyzed data set it has worked relatively fast It works with distances X_{ij}, i,j = 1**, … ,** N, $i \neq j$ between all pairs of the recorded N data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$

2. As an iterative algorithm, it needs some initial ("zero") approximation of the desired projection vectors $\mathbf{y}_1^{(0)}, \ldots, \boldsymbol{y}_N^{(0)}$ It was stated that PCA projections work good as the starting values of CCA; however random choice of initial values works bad: it produces a solution which does not reflect the true shape of the data.

3. There is a substantial difference between the projections obtained by PCA and CCA, the former being of edgy ad the later of rounded shape.

4. Both algorithms permitted to state that the analyzed data set was a mixture of two different classes containing data vectors corresponding to the NO LOAD and LOADED status of the machine.

5. The dydx-plots defined on the base of the CCA method have shown that the NO LOAD component of the mixture constituting the analyzed data has intrinsic dimension much smaller then the other component (see Figure 2).

The novelty in our paper is: (i) The cca method was for the first time applied to gearbox condition monitoring. (ii) We got a better estimation of the intrinsic dimension of the analyzed data. (iii) We indicate clearly that data used usually for gearbox condition monitoring are not Gaussian; moreover, they are not homogenous, thus standard methods used in the methodology of gearbox condition monitoring should undergo a careful revision.

All the results listed in this section apply to data recorded (as vibration sounds) for gearboxes being in a good state, that is, with no (without) serious damage. In our future work we intend to analyze analogous data, however recorded for a device (gearbox) being in bad condition. So far we know, nobody has bothered what is the intrinsic dimensionality of the power spectra data used for example for monitoring gearbox condition.

Acknowledgments

This paper was (in part) financially supported by Polish State Committee for Scientific research 2010-2013 as research project no. N504 147838.

Wydanie publikacji zrealizowano przy udziale środków finansowych otrzymanych z budżetu Województwa Zachodniopomorskiego.

REFERENCES

[1] Bartelmus W., Zimroz R., A new feature for monitoring of planetary gearbox under varying external load, *Mechanical Systems and Signal Processing*, 23/5 (2009) 1528-1534

[2] Bartkowiak A., Zimroz R., Outliers analysis and one class classification approach for planetary gearbox diagnosis, *J. Phys. Conf. Ser.* 305(1) (2011) art. no. 012031, 1-10, IOP Publishing (9th Int. Conf. on Damage Assessment of Structures DAMAS 2011,Oxford UK.)

[3] P. Demartines and J. Herault, Curvilinear component analysis: A self-organizing neural network for non-linear mapping of data sets. *IEEE Trans.on Neural Networks*, 8 (1) (1997) 148-154.

[4] Demartines P., Analyse de donnees par reseaux de neurones auto-organises. PhD thesis. Institut National Polytechnique 1994.

[5] He Q., Yan R., Kong F., Du R., Machine condition monitoring using Principal Component representations, *Mech. Systems and Signal Processing* 23(2) (2009) 446-466.

[6] Herault J., Jausions-Picaud C., Guerin-Dugue A., Curvilinear component analysis for high-dimensional data representation: I. Theoretical aspects and practical use in the presence of noise. In J. Mira and J.V. Sanchez (Eds), *Proceedings of IWANN'99*, vol. II, Springer, Alicante (Spain) June 1999, pp. 625-634.

[7] Trendafilova I., Cartmell M., Ostachowicz W., Vibration based damage detection in an aircraft wing scaled model using principal component analysis and pattern recognition. *Journal of Sound and Vibration* 313 (3-5) (2008) 560-566.

[8] Voiry M., Madani K., Amarger V., Bernier J., Data dimensionality reduction for neural based classification of optical surfaces defects. *International Scientific Journal of Computing (Computing)* vol. 8, issue 1, pp. 32-42.

[9] Zimroz R., Bartkowiak A., Investigation on spectral structure of gearbox vibration signals by principal component analysis for condition monitoring purposes. *J. Phys. Conf. Ser.* 305(1) (2011), art. no. 012075 1-10, (9th Int. Conf. on Damage Assessment of Structures, DAMAS 2011, May 10, Oxford UK).

[10] Zimroz R., Bartkowiak A., Two simple multivariate procedures for monitoring planetary gearboxes in non-stationary operating conditions, *Mechanical Systems and Signal Processing*, DOI: 10.1016/j.ymssp.2012.03.022, in press

[11] Bartkowiak A., Zimroz R., Data dimension reduction and visualization of multidimensional data with application to gearbox diagnostics data: comparison of several methods *Solid State Phenomena* Vol. 180 (2012) 177-184 doi:10.4028/www.scientific.net/SSP.180.177

[12] Osowski, S.; Sikorska-Lukasiewicz, K., PCA transformation and support vector machine for recognition of the noisy images, *Przeglad Elektrot.* 88/3a (2012) 4-6.

***Authors**: dr hab. Anna Bartkowiak, professor emeritus at University of Wroclaw, Institute of Computer Science, Joliot-Curie 15, 50-383 Wroclaw, PL and Wrocław School of Applied Informatics, Wejherowska 28, Wroclaw 54-239, aba@ii.uni.wroc.pl,*

dr hab inż. Radoslaw Zimroz, Diagnostics and Vibro-Acoustics Science Laboratory, Wroclaw University of Technology, Plac Teatralny 2, 50-51 Wroclaw, PL, radoslaw.zimroz@pwr.wroc.pl;

Corresponding author: aba@ii.uni.wroc.pl (A.Bartkowiak)