https://doi.org/10.20965/jaciii.2021.p0090

Paper:

Monitoring and Analysis of Auto Body Precision Based on Big Data

Yixin Yang*,,†, Jianjun Gao**, and Konghui Guo***

*College of Mechanical and Transportation Engineering, Hunan University
Lushan Road, Yuelu District, Changsha, Hunan 410082, China
E-mail: 13913840020@163.com
**BAIC Motor Co., Ltd. Zhuzhou Branch
Liyu Industrial Park, Tianyuan District, Zhuzhou, Hunan 412007, China
†Corresponding author

[Received October 1, 2020; accepted November 17, 2020]

In this paper, a Hadoop-based big data system for auto body precision is established. The system unifies the elements that affect auto body precision into a big data platform, which is more efficient than traditional management methods. Using big data analysis, we devised algorithms to improve the efficiency and accuracy of body precision monitoring. Furthermore, we developed techniques to analyze complex dimension deviation problems using a correlation analysis method, principal component analysis (PCA), and improved PCA method. We further established failure modes and devised monitoring and diagnosis models based on time series analysis.

Keywords: big data, PCA, precision monitoring, intelligent manufacturing, automobile industry

1. Introduction

Dimensional precision is an important indicator of body welding quality. It determines the precision and reliability of the vehicle assembly and further affects the stability and smoothness of the chassis, the handling of the auto, and the matching quality of the interior and exterior trims. However, improvements made on the flexibility of the body production line, which enables the production of mixed lines of multiple models, cause difficulties in debugging and controlling dimensional precision. Thus, the analysis of traditional body precision problems is mostly handled by experienced dimension engineers. However, big data analysis provides a good way to reduce dependence on engineering experience.

The factors that usually influence auto body dimensional precision include the precision of stamping parts, precision and stability of welding fixtures, trolley of automated production lines, precise positioning of the parts shelf, staff operation, and stability of the robot welding system. Thus, multiple factors affect dimensional precision, which are interrelated and influence one another. Hence, it is difficult to analyze and control the causes of deviations.

To investigate this problem, in the 1990s, General Motors, Chrysler, and the University of Michigan jointly launched the "2MM Project" [1]. In China, the study of auto dimensional engineering mainly started when automobile joint ventures adopted technology in their production processes. Systematic research on dimension control in body welding was presented by Lin [2]. Xie et al. also proposed a system optimization method for body parts and fixture design [3]. Yang proposed a method of automatic alarm for body dimension by analyzing the on-line measurement data [4]. Moreover, the concept of a precision quality system for the reference position system of the auto body was proposed by Chen and Huang [5].

As mentioned above, various aspects of auto body dimensional precision control have been investigated in the current literature. Nevertheless, the complexity of the factors that influence body dimensional precision and their multiple sources cause practical issues in dimensional control. The development of intelligent manufacturing and big data technology provides new methods for controlling and improving body precision. In this study, we established a big data system and used big data analysis to monitor auto body dimensional precision and analyze the influencing factors. The relevant measurement information were unified into the big data platform, including 100% online measurement data of more than 1000 key points of the vehicle, regular measurement data of the assembly and sub-assembly parts based on importance, regular and irregular measurement data of more than 500 sets of fixtures, and measurement data of the supplier parts and stamping parts. Afterwards, we automatically analyzed and monitored the auto body precision using a big data analysis model. For multi-deviation source problems, correlation analysis can be used to find the linear correlation between elements. If the deviation source is too complex, principal component analysis (PCA) can be used, whereas for complex non-linear multi-deviation source problems, kernel principal component analysis (KPCA) can be used. The deviation source analysis method based on the time series model is more effective for monitoring recurrent fluctuations.

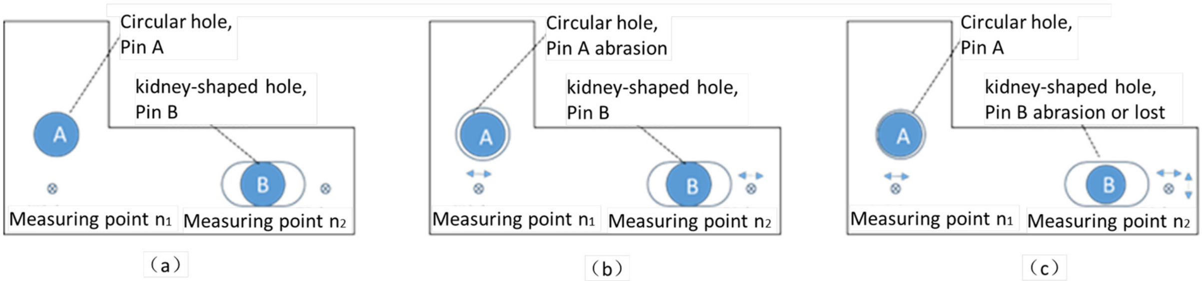

Fig. 1. Parts positioning and wear mode of locating pins.

2. Dimension Correlation Analysis of Multi-Deviation Sources

The change in body dimension generally follows a normal distribution. Hence, dimensional precision control is conventionally based on the principle of normal distribution. The coordinate measuring machine (CMM) method is used to control and improve dimensional precision based on actual measurements. The method is suitable for simple troubleshooting and improvements. However, for multi-factor dimension problems, it cannot effectively determine the causes of deviations and quantitatively analyze the degree of influence of a given deviation source. To address this issue, we introduced a correlation analysis method to quantitatively analyze multiple deviation sources and quickly determine the main factors of the deviation sources.

2.1. Principle of Dimension Correlation Analysis

Correlation coefficient quantifies the degree of correlation between variables. It is a statistical method used to determine if there is a relationship between variables and the degree of such relationship [6]. Covariance is often used to quantify the linear correlation between two random variables. For two variables X and Y, the covariance is defined as [7]:

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}. \qquad (1)$$

Thus, the correlation coefficient γ is defined as:

$$\gamma = \frac{\sum_{i=1}^{n}(x_{1i} - \overline{x_1})(x_{2i} - \overline{x_2})}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \overline{x_1})^2 \sum_{i=1}^{n}(x_{2i} - \overline{x_2})^2}} = \frac{\sum_{i=1}^{n}(x_{1i} - \overline{x_1})(x_{2i} - \overline{x_2})}{(n-1)S_1S_2}, \qquad (2)$$

where x_{1i} and x_{2i} are the i-th measurement values of the first and second variables, respectively, $\overline{x_1}$ and $\overline{x_2}$ are the averages of the first and second variables, respectively, n is the total number of measurements, S_1 and S_2 are the standard deviations of the first and second variables, respectively.

To normalize the correlation coefficient in Eq. (2), we further set $Z_{1i} = (x_{1i} - \overline{x_1})/S_1$ and $Z_{2i} = (x_{2i} - \overline{x_2})/S_2$. Therefore, γ becomes:

$$\gamma = \frac{\sum_{i=1}^{n} Z_{1i}Z_{2i}}{n-1}. \qquad (3)$$

2.2. Dimension Correlation Analysis

As shown in **Fig. 1**, a body part was positioned with circular and kidney-shaped holes on the XY planes. **Fig. 1(a)** shows the positioning of the body parts on the fixture under normal conditions. Circular pins A and B were matched with the circular and kidney-shaped holes, respectively. **Figs. 1(b)** and **(c)** indicate the possible deviation directions of the measuring points n_1 and n_2 under the wear conditions of circular pins A and B, respectively. The failure mode, which causes the size fluctuation of the welding parts, can be determined by performing correlation analysis on the measurement data.

If n_1 and n_2 move horizontally at the same time and with the same magnitude, it indicates that pin A is worn. Similarly, if the movement amplitudes of n_1 and n_2 are not the same, and n_2 moves up and down, it indicates that pin B is smaller or lost. Here, we utilized correlation analysis to determine the type of measurement point fluctuations. In the experiment, CMM was used to measure the dimension. For convenience of measurement, holes were drilled at the n_1 and n_2 positions. The diameters of pins A and B were then measured using a micrometer. The deviation value was obtained by calculating the deviation between the measured and theoretical positions of multiple random clamping measurements.

As shown in **Fig. 1(b)**, pin A is worn and the deviations of n_1 and n_2 are (x_{1i}, y_{1i}) and (x_{2i}, y_{2i}), respectively. The measuring points are listed in **Tables 1** and **2**. The sampling statistics of the correlation analysis was based on more than 20 groups of data randomly selected from hundreds of samples. The correlation coefficients among X1, X2, Y1, and Y2 are:

Table 1. Deviations of the measuring points (n_1, n_2) under the wearing condition of circular pin A. Unit: mm.

	1	2	3	4	5	6	7	8	9	10
X1	0.45	0.35	0.25	−0.3	−0.21	0.11	0.22	0.35	−0.36	−0.23
Y1	0.23	0.12	0.22	0.31	−0.15	−0.13	−0.26	0.18	0.25	−0.21
X2	0.42	0.34	0.24	−0.28	−0.22	0.12	0.23	0.33	−0.35	−0.23
Y2	0.01	0.11	−0.05	0.09	0.07	−0.11	−0.06	0.13	0.05	−0.04

Table 2. Deviations of the measuring points (n_1, n_2) under the wearing condition of circular pin B. Unit: mm.

	1	2	3	4	5	6	7	8	9	10
X1	0.12	0.06	−0.1	0.08	0.02	0.11	−0.11	−0.07	0.09	0.08
Y1	0.06	0.08	0.1	−0.06	−0.05	−0.13	0.07	0.1	0.12	0.05
X2	0.1	0.09	−0.08	0.06	−0.12	0.12	−0.05	0.05	0.05	−0.12
Y2	0.36	0.25	0.45	−0.23	−0.19	0.33	−0.48	0.36	0.25	0.33

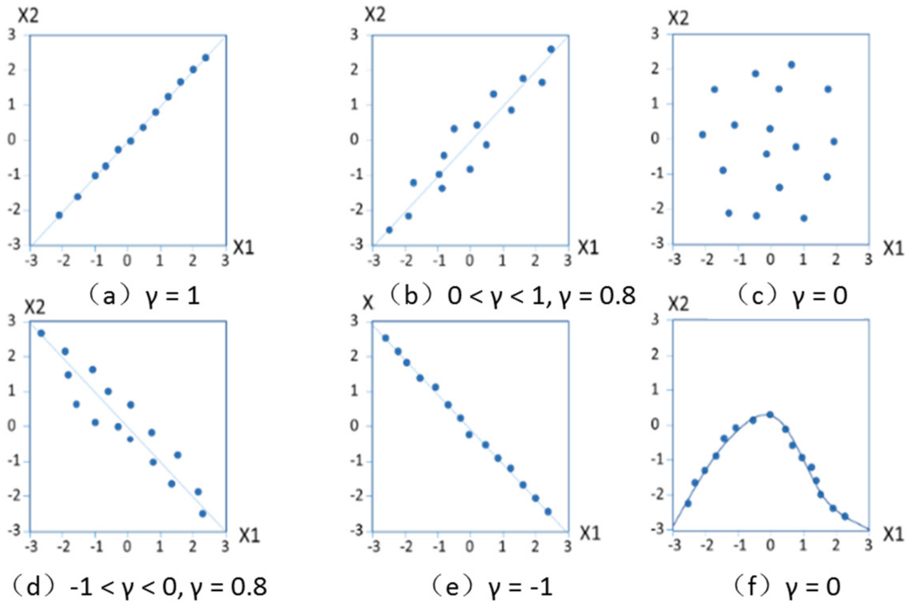

Fig. 2. Scatter diagrams of two variables with different correlation coefficients.

- $\gamma(\mathrm{X1},\mathrm{X2}) = 0.999$, $\gamma(\mathrm{X1},\mathrm{Y1}) = 0.089$,
- $\gamma(\mathrm{X1},\mathrm{Y2}) = -0.055$
- $\gamma(\mathrm{X2},\mathrm{Y1}) = 0.082$, $\gamma(\mathrm{X2},\mathrm{Y2}) = -0.007$,
- $\gamma(\mathrm{Y2},\mathrm{Y1}) = 0.529$

The value of the correlation coefficient indicates the strength of the linear correlation between two variables. As shown in **Fig. 2**, $\gamma = 1$ indicates that the two variables have a clear linear relationship, $0 < \gamma < 1$ indicates that the two variables are positively correlated, $\gamma = 0$ indicates that the two variables do not have a linear relationship, $-1 < \gamma < 0$ indicates that the two variables are negatively correlated, and $\gamma = -1$ indicates that the two variables have a linear relationship with a negative slope.

The correlation coefficient between multiple variables is usually obtained through a computer program. The correlation matrix corresponding to X1, X2, Y1, and Y2 is given in **Table 3**, which shows that the correlation coefficients of X1 and X2 are close to 1. This means that the changes in these two measurements are linearly related. When pin A is worn, n_1 and n_2 synchronously move in the X direction, and the two points have a strong correlation in the X direction. Furthermore, the correlation coefficients in the Y direction are significantly small owing to the limitation of pin B. The movement of n_2 is limited such that there is no obvious correlation between n_1 and n_2. Based on the correlation coefficients and measuring points data, the wear of pin A can be inferred.

For the cases where the correlation coefficient is 0 (see **Fig. 2(f)**), one may conclude that there is no linear relationship between the two variables. However, there is a definite relationship between the two variables, which can be described by a non-linear curve.

In **Fig. 1(c)**, pin B is worn, and points n_1 and n_2 rotate around pin A. The correlation matrix is given in **Table 4**, which shows that the correlation coefficient is small. Moreover, there is no obvious correlation in the XY direction. In this case, correlation analysis cannot ef-

Table 3. Correlation matrix of measuring points n_1 and n_2 under the wearing condition of circular pin A.

Correlation matrix					
		X1	Y1	X2	Y2
Correlation	X1	1.000	0.089	0.999	−0.055
	Y1	0.089	1.000	0.082	0.529
	X2	0.999	0.082	1.000	−0.07
	Y2	−0.055	0.529	−0.07	1.000

Table 4. Correlation matrix of measuring points n_1 and n_2 under wearing condition of round pin B.

Correlation matrix					
		X1	Y1	X2	Y2
Correlation	X1	1.000	−0.412	0.474	0.250
	Y1	−0.412	1.000	−0.141	0.279
	X2	0.474	−0.141	1.000	0.268
	Y2	0.250	0.279	0.268	1.000

fectively indicate the root cause of changes in the measuring point. Hence, a new method needs to be adopted to analyze such changes. Instances of such changes are the rotation mode and multiple parts, multiple processes, and multiple jigs for multiple sources of analysis.

3. Dimension Analysis Based on PCA and KPCA

Correlation coefficient analysis is suitable for determining the linear correlation between multivariate pairs. For a more complex analysis, PCA and KPCA based on kernel function [8] are often used.

3.1. Analysis of Dimension Deviation Sources Based on PCA

Pearson introduced PCA for non-random variables. This concept was further extended by Hotellin to random vectors [9]. PCA is a statistical method for dimensionality reduction. In body dimension analysis, PCA is usually called principal vector analysis. The principal vector are derived and represented by Z_i. Z_i can be interpreted as a linear combination of N original correlation variables X_i as:

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \cdot \\ \cdot \\ \cdot \\ Z_n \end{pmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}. \quad (4)$$

The purpose of principal vector analysis is to determine the eigenvalues and eigenvectors of the correlation matrix. The eigenvalues are the variances of the principal vectors, $\mathrm{Var}(Zi) = \lambda i$. The elements of the eigenvector are $a_{i1}, a_{i2}, \ldots, a_{in}$. The equation that represents the changes

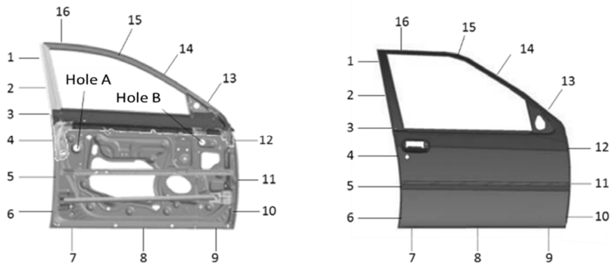

Fig. 3. Measuring points of the door sub-assembly and assembly.

is given by:

$$[\lambda I - C]V = 0, \quad (5)$$

where λ represents the eigenvalue, that is, the variance of the principal vector, I is the identity matrix, C is the covariance matrix, and V is the eigenvector. The eigenvectors characterize the fluctuation via their values and signs of their elements, and they can be used to determine the source of dimension fluctuation and direction. The original n related variables $x_1, x_2, \ldots, x_n$ are used with n master vectors $Z_1, Z_2, \ldots, Z_n$ as:

$$\begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}^{-1} \begin{pmatrix} Z_1 \\ Z_2 \\ \cdot \\ \cdot \\ \cdot \\ Z_n \end{pmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} \begin{pmatrix} Z_1 \\ Z_2 \\ \cdot \\ \cdot \\ \cdot \\ Z_n \end{pmatrix}. \quad (6)$$

Therefore, each original variable is described as a linear combination of multiple principal vectors:

$$x_i = b_{i1}Z_1 + b_{i2}Z_2 + \cdots + b_{in}Z_n. \quad (7)$$

The variance of the original variable is expressed as:

$$V_{ar}[x_i] = b_{i1}^2 v_{ar}[Z_1] + b_{i1}^2 v_{ar}[Z_1] + \cdots + b_{in}^2 v_{ar}[Z_n], \quad (8)$$

$$V_{ar}[x_i] = b_{i1}^2 \lambda_1 + b_{i1}^2 \lambda_2 + \cdots + b_{in}^2 \lambda_n. \quad (9)$$

As mentioned earlier, the correlation matrix is symmetrical, and the number of matrix rows is equal to the number of eigenvalues and eigenvectors.

Here, we used PCA to investigate the causes of deviations to analyze the fluctuations of the door frame size during the welding of the door assembly. As shown in **Fig. 3**, the measuring points 1, 2, 3, 13, 14, 15, and 16 of the door frame significantly fluctuate in the X and Z directions; therefore, it is difficult to find a straightforward rule of the deviation causes. It is also difficult to determine the dimension deviation source of the part's fluctuation using only correlation coefficient method.

To address this issue, we used PCA method to ana-

Table 5. Deviation data of the door frame measuring points. Unit: mm.

	P1			P2			P3			P4			P5			P6			P7		
	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
1	0.60	0.07	0.24	0.35	-.016	0.27	0.14	0.23	0.19	0.03	0.01	1.12	0.15	0.04	0.99	0.13	-0.11	0.50	0.18	-0.13	-0.35
2	0.25	0.28	0.13	0.03	-0.34	0.09	0.08	0.60	0.05	0.18	0.79	0.56	-0.01	0.41	0.50	-0.07	0.62	0.39	-0.01	0.26	0.03
3	0.30	-0.19	0.20	0.10	0.19	0.08	0.08	0.19	0.15	0.17	0.28	0.50	0.19	-0.14	0.40	0.12	-0.13	0.37	0.23	0.17	-0.08
4	-0.50	0.50	-0.07	-0.43	-0.10	-0.19	-0.23	-0.27	0.06	0.18	0.59	-0.70	-0.35	0.33	-0.56	-0.35	0.01	-0.16	-0.36	0.12	0.64
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
20																					

Table 6. PCA results of the deviations of the door frame measuring points.

Component	Characteristic value	Contribution rate	Cumulative [%]
1	11.537	54.939	52.939
2	2.296	10.934	65.873
3	1.902	9.055	74.928
4	1.287	6.128	81.057
5	1.116	5.316	86.373
6	0.933	4.443	90.815
…	…	…	…

Extraction method: Principal component analysis

lyze the above measurement points. The measured data of 20 groups of the selected points are presented in **Table 5** and the PCA results are presented in **Table 6**. Five principal vectors were greater than 1. The characteristic value of the first principal vector was 11.537 and the contribution rate was 54.9%. This explains why the source of the dimension deviation is relatively single.

In the process of applying the PCA method, the data should be analyzed in the XYZ directions. However, because of the high computational complexity involved in determining the main factors, it is more efficient to analyze the measuring points data in the T normal or Y directions rather than in the XYZ directions. As shown in **Fig. 3**, the door fluctuations in the Y direction are unstable. Only the measuring point deviation data in the Y direction and the normal measuring point deviation data orthogonal to the outer plate were considered. We used three methods for measurement and PCA analysis, and the PCA results are presented in **Table 7**.

Table 7 shows that the first three principal components account for 72.4% of the PCA in the XYZ directions. The three principal vectors are relatively dispersed, and in the analysis using the deviations in the Y and T normal directions, the contributions of the first three principal components were 82.9% and 82.3%, respectively. The contribution rate of the first principal vector calculated with the deviation in the T normal direction was 47.3%, which is greater than that of the first principal vector calculated

Table 7. Comparison of PCA results of three data deviation methods (a/b/c).

(a) Explanation of the deviation variance in the Y direction

	Initial eigenvalue		
Component	Characteristic value	Contribution rate	Cumulative [%]
1	6.291	39.321	39.321
2	4.864	30.402	69.724
3	2.107	13.169	82.892
4	1.166	7.286	90.178
5	0.814	5.087	95.265
6	0.295	1.845	97.110
…	…	…	…

(b) Explanation of the deviation variance in the T normal direction

	Initial eigenvalue		
Component	Characteristic value	Contribution rate	Cumulative [%]
1	7.570	47.312	47.312
2	3.667	22.920	70.233
3	1.929	12.057	82.289
4	1.363	8.518	90.808
5	0.604	3.776	94.583
6	0.243	1.519	96.103
…	…	…	…

(c) Explanation of the total PCA variance in the XWZ directions

	Initial eigenvalue		
Component	Characteristic value	Contribution rate	Cumulative [%]
1	5.375	37.092	37.092
2	2.886	19.916	57.008
3	2.237	15.437	72.445
4	1.896	13.084	85.529
5	1.230	8.488	94.017
6	0.243	1.678	95.695
…	…	…	…

Extraction method: Principal component analysis

in the Y direction (39.3%). This indicates that PCA using the deviation in the T normal direction is not only efficient for calculation but also provides a good reference value for determining the principal mode of the deviation source.

3.2. Analysis of Dimension Deviation Sources Based on KPCA Method

Because PCA only decouples data, it cannot be used to analyze nonlinear problems. To address this issue, we utilized a kernel function and applied KPCA to extract the data deviation sources with non-linear features. The KPCA method maps the input space variables from a low-dimensional space to a high-dimensional space through a non-linear function, and then applies PCA on the variables.

In KPCA, the key is to transform the inner product of the characteristic space to the core function of the original space after a nonlinear transformation by introducing the core function. This considerably simplifies the calculation process. Here, we used the core technology of support vector machine to avoid "dimensional disaster." In other words, the inner product operation of samples in the feature space is replaced by a core function that fits the Mercer condition, that is, the core function must be a semi-positive definite function [10].

The corresponding mapping $\Phi : x \rightarrow F$ maps the point x to F by the kernel function. In the corresponding high-dimensional feature space, the variables satisfy the condition of decentralization, that is,

$$\sum_{\mu=1}^{M} \Phi(x_u) = 0. \qquad (10)$$

The covariance matrix of the feature space is then given by:

$$C = \frac{1}{M} \sum_{u=1}^{M} \Phi(x_u) \Phi(x_u)^T, \qquad (11)$$

and the eigenvalues and eigenvectors of C are obtained using:

$$V \varepsilon F \backslash \{0\}, \quad Cv = \lambda v. \qquad (12)$$

The eigenvector can be expressed as a linear combination of $\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_M)$, where $v = 1, 2, \ldots, M$. The eigenvalues and eigenvectors are obtained by solving the $M \times M$ matrix K. Thus, the projection of the test variables in the new eigenvector space v^k is given by:

$$\left(v^k \cdot \Phi(x)\right) = \sum_{i=1}^{M} (\alpha_i)^k \left(\Phi(x_i), \Phi(x)\right). \qquad (13)$$

Replacing the inner product with the kernel function,

$$\left(v^k \cdot \Phi(x)\right) = \sum_{i=1}^{M} (\alpha_i)^k K(x_i, x). \qquad (14)$$

If Eq. (14) is not established, then it is adjusted as follows:

$$\Phi(x_u) \rightarrow \Phi(x_u) - \frac{1}{M} \sum_{i=1}^{M} \Phi(x_v), \quad u = 1, 2, \ldots, M. \qquad (15)$$

The kernel matrix is then modified as:

$$K_{uv} \rightarrow K_{uv} - \frac{1}{M} \left(\sum_{w=1}^{M} K_{uw} + \sum_{w=1}^{M} K_{wv} \right) + \frac{1}{M^2} \sum_{w,\tau=1}^{M} K_{w\tau}. \qquad (16)$$

Based on the principle of KPCA, the relevant calculation process is as follows.

1. Express the n indexes obtained in a $m \times n$-dimensional matrix (assuming each index has m samples).
$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \ldots & \ddots & \ldots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \qquad (17)$$
2. Standardize the matrix and set $X = (x_{ij})_{m \times n}$.
3. Calculate the correlation matrix, $R = 1/(m-1) X^T \cdot X = (r_{ij})_{n \times n}$.
4. Obtain the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and the corresponding eigenvectors $v_1, v_2, \ldots, v_n$ using the Jacobian iterative method.
5. Sort the strong eigenvalues in descending order to obtain $\lambda_1' > \lambda_2' > \cdots > \lambda_n'$ and the corresponding adjusted eigenvectors $v_1', v_2', \ldots, v_n'$.
6. Perform unit orthogonalization on the eigenvectors using Schmidt orthogonalization to obtain $\alpha_1, \alpha_2, \ldots, \alpha_n$.
7. Calculate $\{B_1, B_2, \ldots, B_n\}$ to obtain the cumulative contribution rate of the eigenvalues. Afterwards, set the value of the extraction efficiency ρ. If $B_t \geq \rho$, then t principal components $\alpha_1, \alpha_2, \ldots, \alpha_t$ are extracted.
8. Calculate the projection $Y = X \cdot \alpha$ of the sample variable (standardized X) on the extracted eigenvectors, where $\alpha = \alpha_1, \alpha_2, \ldots, \alpha_n$. Y is the dimension-reduced data.

The advantage of KPCA over PCA is that the latter is an algebraic feature analysis method. PCA requires a large memory space, and its algorithm is computationally complex. For an original space with dimension n, PCA needs to decompose a $n \times n$ non-sparse matrix. Furthermore, as a linear mapping method, the dimension-reduced expression of PCA is generated by linear mapping. Therefore, it ignores non-linear relationships in data samples and overlooks the optimal features. This is the main reason why the PCA method is not effective in some cases [11]. However, KPCA uses a non-linear method to extract the principal components, which map the variables to a high-dimensional space F through a mapping function, and subsequently uses PCA to analyze the function space F.

Here, we consider the deviation source analysis of the body side tail-light installation area as an example to compare the analysis results of the PCA and KPCA methods. The dimension fluctuation in the mounting area of the side tail light is an issue in body dimension control (see **Fig. 4**) because the matching relationship between this area and the tail lights, rear bumper, and rear trunk is complicated, and the dimension of the rear body side area is the key to the matching quality of this area.

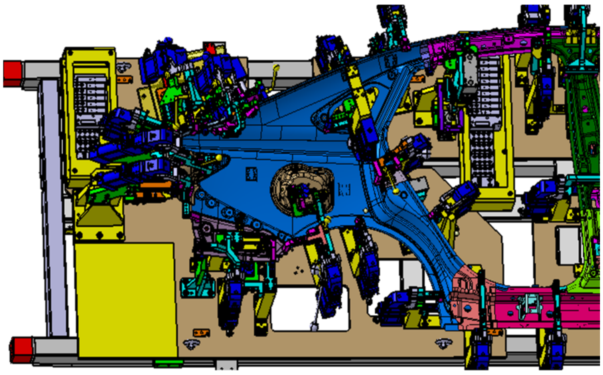

Fig. 4. Clamping drawing of the rear area of the body side.

We used both PCA and KPCA methods to analyze the measurement data and then compared the results. Several commonly used kernel functions are linear kernel function, P-order polynomial kernel function, Gaussian radial basis function kernel function, and multilayer perceptron kernel function. Considering the characteristics of the analysis of the deviation source of the flexible sheet, we adopted a polynomial kernel function given by:

$$K(x \cdot x_i) = [(x \cdot x_i) + 1]^p. \qquad (18)$$

Table 8 presents the comparison of the analysis results of PCA and KPCA. The contribution rate of the third principal component obtained using PCA was 63.9%, whereas that of the first principal component was 26.9%. Meanwhile, the contribution rate of the first three principal components obtained using KPCA was 80.7%, whereas that of the first principal component was 62.5%. Thus, the contribution rate of the first principal component of KPCA is higher than that of PCA. It is easier to determine the main mode of deviation. According to the on-site analysis, due to the loose clamps of the body side tail and the rebound of the outer plate of the body side tail, the rear light assembly area is rotated around a certain axis. After the clump is repaired, the measuring points in this area become stable and qualified.

4. Analysis Based on the Model of Dimension Deviation Source

The deviations of part assemblies are often caused by the deviations of sub-assemblies or single parts. In such cases, we analyzed the data to identify the deviation sources. The data here refers to the corresponding measuring points of the assembly, sub-assembly, or single parts. We studied two analysis methods to identify the deviation source: one is based on the failure mode of the parts and fixtures, and the other is based on time series.

4.1. Principal Vector Analysis Based on Failure Mode

PCA is used to extract the principal vectors (deviation mode vectors) p_i and s_j of parts or sub-assemblies and assemblies. The degree of vector correlation is expressed by the correlation coefficient as:

$$\eta_{ij} = \left| \frac{p_i \cdot s_j}{|p_i||s_j|} \right|, \qquad (19)$$

where η_{ij} is the correlation coefficient between p_i and s_j, p_i is the i-th principal component of the part or sub-assembly p, and s_j is the j-th principal component of the assembly s. To eliminate the noise interference as much as possible, we set a threshold V_{comp} and formed a mapping relationship between part p and assembly s when the correlation coefficient was greater than V_{comp}. The contribution rate μ represents the strength of the mapping relationship, and it is given by:

$$\mu = \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{m=1}^{m} \lambda_m}, \qquad (20)$$

where λ_j is the eigenvalue of s_j corresponding to p_i, q is the number of eigenvectors greater than V_{comp} (if $q \neq 1$), and m is the total number of extracted eigenvectors. The threshold V_m of the first feature principal vector is set to reduce the computational complexity of the correlation analysis. If the contribution of the first principal component extracted by a part is less than V_m, it suggests that the part has no significant impact on the deviation of the assembly. However, if the contribution of the first principal component is greater than V_m, it is considered to have a significant influence on the deviation of the assembly [12].

In addition to the deviation source of the welding assembly caused by the parts, another important factor is the fixtures. The analysis of the deviation source generated by the fixtures is based on investigating the potential failure modes of the direction and the amplitude of the movement at each measuring point.

According to the 3-2-1 positioning principle for parts, the failure modes of fixtures are divided into failure of the main positioning pin, failure of the secondary positioning pin, and failure of the positioning and clamping points. The corresponding parts also appear in three modes: translation along the AB direction, rotation in the plane around point A, and rotation in the space around the axis formed by AB. The failure modes are shown in **Fig. 5**.

For the first failure mode, most of the measuring points in **Fig. 5(a)** move along the kidney-shaped hole and slightly rotate around B. The deviation mainly appears in a specific direction, and the deviation is the same. The failure mode in this case can be obtained by unitizing the column vector comprising all the measuring points. In **Figs. 5(b)** and **(c)**, the measuring point deviation mainly

Table 8. Comparison of PCA and KPCA analysis results of rear body side deviation.

PCA				KPCA			
Component	Characteristic value	Contribution rate	Cumulative [%]	Component	Characteristic value	Contribution rate	Cumulative [%]
1	2.536	26.921	26.921	1	0.285	62.500	62.500
2	1.809	19.024	46.125	2	0.062	13.596	76.096
3	1.678	17.813	63.938	3	0.021	4.605	80.702
4	1.106	11.741	75.679	4	0.017	3.728	84.430
5	0.812	8.620	84.299	5	0.012	2.632	87.061
6	0.678	7.197	91.497	6	0.008	1.754	88.816
…	…	…	…	…	…	…	…

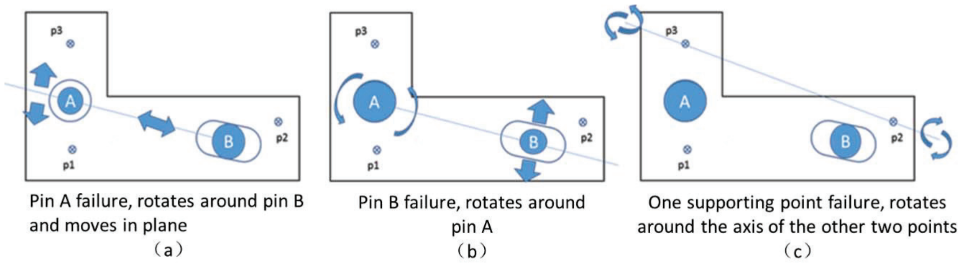

Fig. 5. Three failure modes of fixtures.

has a fixed axis for rotating motion. Suppose the vector of the rotation direction of the measuring point is $e_i = (x_i', y_i', z)$, the distance from the measuring point to the center of rotation is d_i' and $e_i \times d_i'$, and the column vector is $[e_1 \cdot d_1, e_2 \cdot d_2, \ldots, e_n \cdot d_n]^T$. After unitization, $e_i \times d_i'$ is the failure mode under rotation.

For the fixture failure mode, the analysis of the deviation source of the welding assembly is based on calculating the correlation coefficient η_{ik} between the fixture failure mode a_i and the main eigenvector v_k of the assembly measurement point data as follows:

$$\eta_{ik} = \left| \frac{a_i \cdot s_j}{|a_i||s_j|} \right|. \quad (21)$$

To reduce the amount of calculation, a threshold value V_{jig} was considered, and a mapping relationship was determined between all the main eigenvectors larger than V_{jig} and the failure modes. To determine the contribution degree of the main eigenvector in the assembly deviation, the deviation contribution coefficient ω is defined as:

$$\omega_i = \eta_{ik} \cdot \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \cdot 100\%, \quad (22)$$

where ω_i represents the deviation contribution factor of the i-th failure mode of the fixture, λ_k is the corresponding eigenvalue of v_k, and p is the number of main eigenvectors of the welding assembly. By comparing the contribution coefficients of different deviation modes, the failure modes and points of the locations can be analyzed and checked.

4.2. Discrimination of Dimensional Deviation Source Based on Time Series Analysis

Here, we first established a time series autoregressive (AR) model to process the continuous measurement data and obtain a stable normal distribution with zero mean. The first step is to extract the trend term to measure the time series data $\{x_t\}$ $(t = 1, 2, \ldots, N)$, remove the non-stationary part $d_t : y_t = x_t - d_t$, and form stationary time series $\{y_t\}$. Afterwards, the time series are zeroed and normalized to obtain the normal distribution $\{x_t\}$, where $x_t \sim N(0, 1)$. The basic expression of AR(n) is:

$$\begin{aligned} x_t &= \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \cdots + \varphi_n x_{t-n} \\ &\quad + a_t a_t \sim NID\left(0, \sigma_a^2\right), \end{aligned} \quad (23)$$

where $a_t - \{x_t\}$ $(t = 1, 2, \ldots, N)$ residuals correspond to a normal distribution with variance σ^2, $\varphi_1, \varphi_2, \ldots, \varphi_n$ and σ_a^2 are the parameters $\{x_t\}$ $(t = 1, 2, \ldots, N)$ estimated by a time series method. Generally, the least-squares method is used for parameter estimation, which is relatively simple and unbiased. The least-square estimation of the model is calculated by:

$$\hat{\varphi} = \left(x^T x\right)^{-1} x^T y, \quad (24)$$

where $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_n)^T$, $y = [x_{n+1}, x_{n+2}, \ldots, x_N]^T$, and

$$x = \begin{bmatrix} x_n & x_{n-1} & \cdots & x_1 \\ x_{n+1} & x_n & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{N-1} & x_{N-2} & \cdots & x_{N-n} \end{bmatrix}.$$

PCA is used to extract the feature roots $\lambda_1, \lambda_2, \ldots, \lambda_m$ and eigenvector V from an n-dimensional pattern vector $\varphi =$

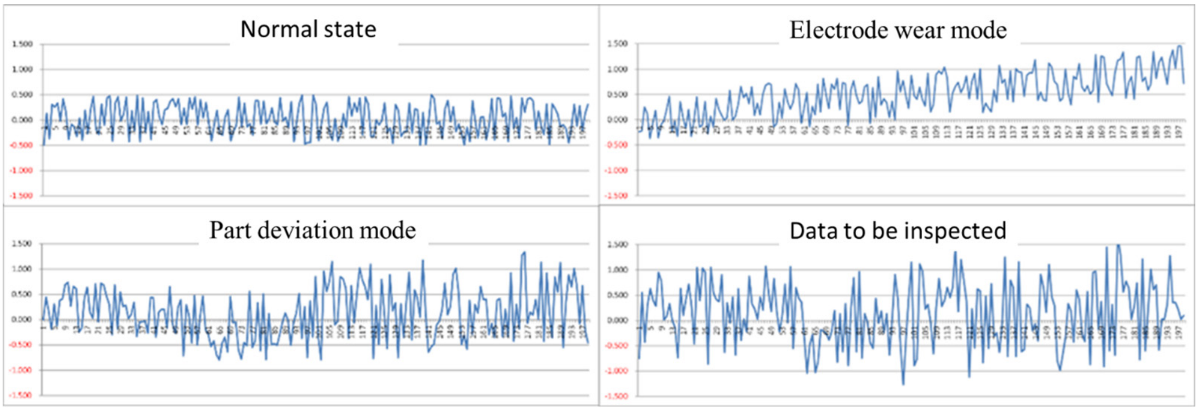

Fig. 6. Deviation data of the body top cover measurement points.

$(\varphi_1, \varphi_2, \ldots, \varphi_n)^T$. The vector V is a low-dimensional eigenvector obtained by the PCA reduction of the pattern vector φ as:

$$V_{m\times 1} = A_{m\times 1}\varphi_{n\times 1}. \qquad (25)$$

To construct the discriminant function for a certain part, corresponding K reference states are obtained for the K dimension deviation sources $F_{R1}, F_{R2}, \ldots, F_{RK}$. The constructed function is a discriminant function. Then classify the part's mode vector φ_t, analyze, and judge its deviation mode.

The discriminant function adopts the distance discriminant function, and its measurement method is the geometric Euclidian distance. The geometric distance between φ_t and F_R in an n-dimensional geometric space is expressed as the sum of squares of the coordinate difference of two points in space:

$$D^2(X,Y) = \sum_{i=1}^{n}(x_i - y_i)^2 = (X,Y)^T(X,Y), \qquad (26)$$

where $X = [x_1 x_2 \cdots x_N]^T$ and $Y = [y_1 y_2 \cdots y_N]^T$ are arbitrary points in space. The time series data $\{x_t\}_T$ to be checked is transformed into a coefficient matrix X_T, which is substituted in the AR model to form the reference deviation:

$$X_T\emptyset_R = a_{RT}X_T\emptyset_{RT} = a_T, \qquad (27)$$

where a_{RT} is the calculated residual vector of the coefficient matrix XT to be tested, $\emptyset_R$ is the reference model parameter, $X_T\emptyset_{RT} = a_T$ is the AR model to be tested. The geometric distance between a_T and a_{RT} represents the residual offset distance $D^2(a_T, a_{RT})$, given by:

$$D^2(a_T, a_{RT}) = N_T(\emptyset_T - \emptyset_R)^T R_T(\emptyset_T - \emptyset_R), \qquad (28)$$

where R_T is the covariance matrix of the test time series data $\{x_t\}_T$ and N_T is the length of the test time series data $\{x_t\}_T$. Let N_T be equal to the length of the reference deviation mode time series data, that is, $N_T = N_R = N$.

It can be seen from Eq. (28) that $D^2(a_T, a_{RT})$ is a function of $\emptyset_T$ and $\emptyset_R$, which can be expressed as a weighted residual offset distance measured in $n+1$ dimensional space. The offset distance function related to the residual φ is then derived as:

$$D^2(a_T, a_{RT}) = N_T(\varphi_T - \varphi_R)^T r_T(\varphi_T - \varphi_R), \qquad (29)$$

where r_T is the covariance matrix of the time sequence to be checked $\{x_t\}_T$, which is equivalent to the n-th-order sub-matrix of R_T minus the first row and first column. For the discrimination of K deviation source test modes, we selected the reference mode with the smallest $D^2(a_T, a_{RT})$ value. Thus,

$$D_a^2(\varphi_T, \varphi_{R(j)}) = \min\left\{D_a^2(\varphi_T, \varphi_{R(i)})(i = 1, 2, \ldots, K)\right\}\varphi_T \in F_{Rj}. \qquad (30)$$

The reference population where φ_T is located should satisfy the condition of minimum residual offset distance.

The geometric distance calculation and deviation diagnosis are as follows: First, the deviation source is classified. According to the corresponding deviation state of the dimension deviation data, an AR model is established to obtain the mode vector. Afterwards, the principal vector is extracted to obtain the covariance. Using Eqs. (29) and (30), the distance is evaluated and the corresponding deviation mode is established for reference. The next step is to model the inspection data. The geometric distances under different deviation states are also obtained and compared to determine the dimension deviation source [13].

Here, we consider the upper edge of the windshield glass of the auto body roof. The normal measurement point matching the glass often fluctuates mainly due to the wear of the welding electrode cap or the deformation of the incoming material. Two deviation modes corresponding to the wear of the electrode cap and the deformation of the incoming material can be established. We collected real-time data of the implementation detection data. By calculating, tracking, and comparing the pre-established deviation modes, we automatically identified the deviation modes that eliminate the need for manual inspection.

The data under three states (normal operation, incoming material deformation, and electrode cap wear) were collected and shown in **Fig. 6**. The data to be inspected is a group of actual measurement data of parts with dimension fluctuations during welding. Two hundred groups

Table 9. Euclidean distance and squared Euclidean distance.

	Approximation matrix Euclidean distance				Approximation matrix squared Euclidean distance			
	Normal deviation mode	Part deviation mode	Electrode wear mode	Actual test data	Normal deviation mode	Part deviation mode	Electrode wear mode	Actual test data
Normal deviation mode	0.000	0.300	0.370	0.223	0.000	0.090	0.137	0.050
Part deviation mode	0.300	0.000	0.234	0.083	0.090	0.000	0.055	0.007
Electrode wear mode	0.370	0.234	0.000	0.218	0.137	0.055	0.000	0.047
Actual test data	0.223	0.083	0.218	0.000	0.050	0.007	0.047	0.000

of time series data were selected, and the total time series span was three days. The time series model was then established according to the AR modeling requirements, and the feature vector was extracted using the PCA method. The AR model under three states is established as follows:

1. Normal:
$$x_t = -0.567x_{t-1} + 0.112x_{t-2} - 0.093x_{t-3} + 0.112x_{t-4} - 0.101x_{t-5} + a_t$$
2. Incoming material deformation:
$$x_t = -0.583x_{t-1} + 0.159x_{t-2} - 0.065x_{t-3} - 0.063x_{t-4} + 0.136x_{t-5} + a_t$$
3. Electrode cap wear:
$$x_t = -0.505x_{t-1} + 0.012x_{t-2} + 0.062x_{t-3} - 0.148x_{t-4} + 0.077x_{t-5} + a_t$$

Based on the measured data of the workshop, the AR model is established as follows:

$$x_t = -0.579x_{t-1} + 0.131x_{t-2} - 0.053x_{t-3} - 0.028x_{t-4} + 0.067x_{t-5} + a_t.$$

The Euclidean distance and squared Euclidean distance between the inspected data and the data of other modes are presented in **Table 9**, in which the deviation mode between the measured data and the reference sample is the smallest. This indicates that the deviation mode might be the same or quite similar. Therefore, the deviation mode of the measured data is most likely the deviation mode caused by the part fluctuation. Another distance calculation method is the Euro-square distance, whose calculation results are 0.505 (normal mode), 0.007 (part deviation mode), and 0.047 (electrode wear mode). It can be seen from the data that the squared Euclidean distance provides a better and more deviation mode discrimination.

5. Conclusion

1. The application of big data analysis improved the utilization efficiency of measurement resources and detection data of auto body precision systems, and also reduced the dependence on the experience of engineers. It is more efficient than traditional methods that rely on an engineer's experience. Big data analysis facilitates the analysis of the precision of body dimensions.
2. Correlation analysis, PCA, and KPCA analysis methods are capable of diagnosing complex dimension deviation problems. The failure-mode-based principal component analysis model and dimension deviation source monitoring diagnosis model based on time series analysis were also shown to effectively improve the efficiency of failure modes and deviation source diagnosis.

Acknowledgements

The data used to support the findings in this paper are available upon request from the corresponding author. The authors declare that there are no conflicts of interest regarding the publication of this paper. This study was supported by the China State Key Laboratory of Advanced Design and Manufacturing of Automobile Body Open Fund Project (No. 31515010).

References:

[1] D.-H. Tang, "From American Automobile 2mm Program to China Automobile Manufacturing Quality Program," Development & Innovation of Machinery & Electrical Products, Vol.2006, No.1, pp. 26-28, 2006 (in Chinese).

[2] Z. Lin, "Quality control technology of automobile body manufacturing," China Machine Press, 2005.

[3] H. Xie, C. Jiang, and Z. Zhang, "Interval description of dimensional tolerance and uncertainty optimization," The 3rd Conf. on Computational Mechanics (CCM 2014), 2014.

[4] D. Yang, "Research on real-time monitoring and fault diagnosis system and its key technologies in BIW welding process," Master Thesis, Hefei Polytechnic University, 2014.

[5] X. Chen and J. Huang, "Application of RPS in Precision Design Body," Automotive Technology, Vol.2006, No.8, pp. 18-21+42, 2006 (in Chinese).

[6] Y. Tao, H. Shi, B. Song, and S. Tan, "A Novel Dynamic Weight Principal Component Analysis Method and Hierarchical Monitoring Strategy for Process Fault Detection and Diagnosis," IEEE Trans. on Industrial Electronics, Vol.67, No.9, pp. 7994-8004, 2020.

[7] T. Sibillano, A. Ancona, V. Berardi, and P. M. Lugarà, "Correlation analysis in laser welding plasma," Optics Communications, Vol.251, Issues 1-3, pp. 139-148, 2005.

[8] S. Yi, Z. He, X.-Y. Jing, Y. Li, Y.-M. Cheung, and F. Nie, "Adaptive Weighted Sparse Principal Component Analysis for Robust Unsupervised Feature Selection," IEEE Trans. on Neural Networks and Learning Systems, Vol.31, No.6, pp. 2153-2163, 2020.

[9] H. Caussinus and A. Ruiz-Gazen, "Principal Component Analysis, Generalized," S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, and N. L. Johnson (Eds.), "Encyclopedia of Statistical Sciences," Vol.10, John Wiley & Sons, Inc., 2005.

[10] Z. Sun, J. Zhao, and L. Li, "Comparison between PCA and KPCA methods in comprehensive evaluation of robotic kinematic dexterity," High Technology Letters, Vol.20, No.2, pp. 154-160, 2014.

[11] S. Yang, "Research on BIW size deviation based on wavelet modulus maximum and principal component analysis," Master Thesis, Chongqing Jiaotong University, 2014.

[12] Y. Yang, S. Shen, Y. Liu, Z. Li, and S. Jin, "Diagnosis of body online detecting dimensional deviation based on pattern recognition," Machinery Design & Manufacture, Vol.2012, No.1, pp. 188-190, 2012 (in Chinese).

[13] K. Geng, J. He, K. Wang, and Y. Chen, "Gear Fault Diagnosis Based on Pattern Recognition of Acoustic Emission Signals," Machine Tool & Hydraulics, Vol.47, No.16, pp. 192-196+208, 2019 (in Chinese).

Name:
Yixin Yang

Affiliation:
Hunan University

Address:
Xianshanmeijing Tianyuan District, Zhuzhou, China

Brief Biographical History:
1997- College of Mechanical and Automotive Engineering, Hunan University
2011- Visiting Researcher, University of Warwick
2014- Doctor, Hunan University

Main Works:
• Y. Yang et al., "Construction and Implementation of Intelligent Transparent Automobile Factories," China Mechanical Engineering, Vol.23, No.29, pp. 2867-2874, 2018.
• Y. Yang et al., "Technology of robot dormancy based on industrial internet control," J. of Machine Design, Vol.37, No.3, pp. 14-18, 2020.

Membership in Academic Societies:
• Zhuzhou Association for Science and Technology, Vice Chairman
• Technical Committee of Zhuzhou Automobile Industry Association, Deputy Director

Name:
Jianjun Gao

Affiliation:
BAIC Motor Co., Ltd. Zhuzhou Branch

Address:
Liyu Industrial Park, Tianyuan District, Zhuzhou, Hunan 412007, China

Brief Biographical History:
1992- Jilin University of Technology
2017- President, BAIC Motor Co., Ltd. Zhuzhou Branch

Main Works:
• J. Gao and Y. Zhao, "Planning, design and application of auto transfer centering mechanism for car conveyor line," Enterprise Technology Development, 2014.
• Y. Yang, J. Gao, Z. Pi, and K. Guo, "Construction and Application of Big Data System in Automobile Intelligent Factory," 2nd Int. Conf. on Artificial Intelligence and Computer Science (AICS2020), 2020.

Membership in Academic Societies:
• Zhuzhou Automobile Industry Association, President

Name:
Konghui Guo

Affiliation:
Hunan University

Address:
Lushannan Road, Yuelu District, Changsha, China

Brief Biographical History:
1991- Awarded the title of doctoral supervisor
1993- Vice President, Jilin University of Technology
1994- Member of the Chinese Academy of Engineering

Main Works:
• His research projects computer dynamic simulation of vehicle handling and stability, tire mechanical proper-ties, vehicle cornering and braking stability, and dynamic simulation of man vehicle closed-loop control system, and research on several national standards successively won ministerial and national science and technology progress awards of China.

Membership in Academic Societies:
• Chinese Academy of Engineering