Optimizing precision of SIFT algorithm in feature extraction of tennis video

Changhong Wu[1], Haiyan Geng[1*], Tianlin Geng[2]

[1]*College of Sports, Langfang Teachers College, Langfang, 065000, Hebei, China*

[2]*Baoding Gaoyang Hongrun Middle School, Baoding, 071500, Hebei, China*

Received 1 June 2014, www.cmnt.lv

Abstract

The traditional SIFT algorithm still has problems such as running slowly and low accuracy in the tennis video feature extraction and matching, an improved SIFT algorithm is proposed based on a tennis video feature extraction and matching. First, it limits the number of feature points to SIFT algorithm by adding the image texture features, which make the feature points to evenly distribute in each set of different scales of video image. Then measures the similarity of feature points by Euclidean distance, the measurement results transform with projection transformation relations, and then uses iterative arithmetic of random sampling consistency (RANSAC) algorithm to obtain maximum satisfy feature points of geometry model. Finally, uses the minimum root mean square error (RMSE) to determine the accuracy of registration. The simulation experiments show that the proposed improved SIFT algorithm based on tennis video feature extraction and matching has faster running speed and better matching precision.

Keywords: improved SIFT algorithm, tennis video, feature extraction, feature matching, the texture features, projection transformation, random sampling consensus

1 Introduction

With the advance of science and technology in recent years, especially the development and promotion of digital technology, storage cost is reduced; the growth of the network bandwidth and the processing speed of computer are improved, the digital video information is in the rapid expansion [1]. Sports video content analysis technology as a separate issue to the attention of the researcher, has a broad application of prospects, and also has an important academic value. The core technology of sports video content analysis is actually on the analysis of the events and their relationships [2]. It can meet the demand of most of the users by the study of the annotation and organization of these events. From the view of academic, incident detection and recognition is a typical problem in computer vision, pattern recognition.

The analyses of current sports video usually use three layers framework, namely the low-level feature layer, intermediate object layer and senior events layer [3]. Lew et al. discuss the goal of a snooker game testing method, using color information segment the ball and the ball bag, and track the ball near the ball bag, thus to detect the goal events [4]. Li and others take a statistics method based on rules method and hidden markov model, by the analysis of characteristics such as the ground color and athletes' clothing color, classify the football match and suspended scene[5]. Smith and Chang extract statistics (mean value and variance) in small wave band as texture, good results have been achieved [6]. Ma and Manjunath evaluat various wavelets transform, including orthogonal and biorthogonal wavelet transform, tree structure wavelet transform and Gabor wavelet transform. They found that the Gabor wavelet transform is the most accord with human vision characteristics [7]. Huang et al. achieve automatic learning and testing of the replay scene by the logo process of describing replay scene with motion characteristics, good effects have been obtained [8]. two new spectrum characteristics for audio classification is proposed by Erwin et al., it extracts audio signal characteristics, and then classifies the test data according to the threshold value or does event detection [9]. Assfalg et al. found that space camera is usually characterized by the unity of the large color area and the regular ground line, the individual is more prominent background is fuzzy for the athletes close-up, and viewers messy performance in scenes, individual is not clear, because of these problems, they identify the three types of lens effectively by neural network with extracting the texture, edge and color features [10]. Duan et al. combine the corresponding domain knowledge according to field area, the field lines and ground object scale characteristics, using decision tree to program scene classification such as football, basketball, tennis, volleyball and so on [11]. Stauffer and Grims are modeling for each pixel in the scene by using gaussian mixture model, and real-time update model by online approximation estimation method, match the current of each pixel in the image with its corresponding matching gaussian mixture model, divide all pixels into foreground pixels and background pixels according to the matching

[*] ***Corresponding author's*** e-mail: wch-9140@163.com

conditions [12]. Saurt et al. realized the automatic analysis of basketball video content by directly using MPEG compressed domain feature, algorithm to detect specific events, such as stealing, rushing and possible shooting, etc. mainly based on the statistical analysis of motion vector size and direction [13]. A framework based on knowledge of semantic reasoning is proposed by Wu et al. to identify the sports event in the video, it includes three layers, the lower extracts features and segments video. Middle layer gives semantic for the segmentation of video clips concepts by using neural network and decision tree. Finally, high layer reasons to identify events based on already defined finite automata model [14]. Ling Yu also describes video of high layer semantic information by the middle layer, acquires the low layer features, and then forms the middle layer by using some clustering method and machine learning algorithm, finally, makes use of sports video proprietary rules to get top events [15].

As the traditional SIFT algorithm still has problems such as running slowly and low accuracy in the tennis video feature extraction and matching, we propose an improved SIFT algorithm, which optimizing the feature extraction and matching process of the feature point distribution of traditional algorithm.

2 The defects analysis of SIFT algorithm

SIFT algorithm is a method of feature matching with widely used in computer vision field. It can extract the feature information of the object from the image, these feature points for the scale of the image scaling and rotation, a degree of light intensity and camera Angle changes has invariance. These feature points can effectively decrease the destruction of object structure caused by the screening, chaos and noise. In addition, the feature points with high specificity extracted by the SIFT algorithm, ensures the precision of matching algorithm.
With the advancement in networking and multimedia technologies enables the distribution and sharing of multimedia content widely. In the meantime, piracy becomes increasingly rampant as the customers can easily duplicate and redistribute the received multimedia content to a large audience.

2.1 THE EXTREME DETECTION OF SCALE SPACE

The first step of key point detection is for pixel location which not sensitive with image scale changes on the scales. Detection on different scale is for analysis to the same object from different perspectives. This process can use the continuous function in the scale space to calculate. Gaussian function has been shown to be the only possible kernel function under a series of reasonable assumptions. So space scale of the image is defined as a function $L(x, y, \sigma)$, it gets by the convolution of the input image $I(x, y)$ and the different scales of Gaussian kernel function $G(x, y, \sigma)$:

$$L(x, y, \sigma) = G(x, y, \sigma) \cdot I(x, y), \tag{1}$$

where "$\cdot$" is convolution for x and y.

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{x^2+y^2}{2\sigma^2}}, \tag{2}$$

In order to effectively detect point position in image scale space, it uses Gaussian differential function $D(x, y, \sigma)$, which can be calculated by subtracting image convolution results of two adjacent the scales:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, (k-1)\sigma)) \cdot I(x, y) = L(x, y, k\sigma) - L(x, y, (k-1)\sigma), \tag{3}$$

The relation between Gaussian differential function and LOG Laplacian of Gaussian $\sigma^2\nabla^2 G$ can analyze by caloric equation.

A function $u(x, y, z, t)$ to 3D space coordinates (x, y, z) and time variable t, caloric equation is as follows:

$$\frac{\partial u}{\partial t} - k\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right) = 0. \tag{4}$$

Or be equal:

$$\frac{\partial u}{\partial t} = k\nabla^2 u, \tag{5}$$

where, k is a constant in equation.

According to the second kind of caloric equation, we can get the following equation:

$$\frac{\partial G}{\partial \sigma} = \sigma\nabla^2 G. \tag{6}$$

It can be seen from above equation, $\sigma\nabla^2 G$ can express by finite differential form of $\frac{\partial G}{\partial \sigma}$:

$$\sigma\nabla^2 G = \frac{\partial G}{\partial \sigma} \approx \frac{G(x, y, k\sigma) - G(x, y, (k-1)\sigma)}{\sigma}. \tag{7}$$

Then:

$$G(x, y, k\sigma) - G(x, y, (k-1)\sigma) \approx \sigma^2\nabla^2 G. \tag{8}$$

It show that Gaussian differential function is the approximate form of LOG Laplacian of Gaussian $\sigma^2\nabla^2 G$.

2.2 ACCURATE POSITIONING OF KEY POINTS

The extremum points in step 1 as alternative point use Taylor expansion of scale space function $D(x, y, \sigma)$:

$$D(X) = D + \frac{\partial D^T}{\partial X} X + \frac{1}{2} X^T \frac{\partial^2 D}{\partial X^2} X, \tag{9}$$

where $D(X)$ and its expansion term were calculated in the sampling point, $X=(x,y,\sigma)^T$ is the compensation dosage for that point. When the expansion term of $D(X)$ is 0, the corresponding X as an extreme point, its position determine by the following equation:

$$\hat{X}=-\frac{\partial^2 D^{-1}}{\partial X^2}\frac{\partial D}{\partial X}. \tag{10}$$

The difference value between adjacent sampling points in specific calculation is used to approximately describe the scale space function $D(x,y,\sigma)$ and its expansion term. If the compensation dosage $\hat{X}$ is greater than 0.5 in various scales, then the distance between the current sample point and the extremum point is not the shortest. Next out the current sample point, to continue calculate the same for the other sample points. Finally get the approximate estimates of extreme value point location by together $\hat{X}$ and its corresponding sampling points.

The value $D(\hat{X})$ of the scale space function in the extreme value point will be used to eliminate the low contrast and unstable extreme value point. Substituting Equation (10) into Equation (9):

$$D(\hat{X})=D+\frac{1}{2}\frac{\partial D^T}{\partial X}\hat{X}. \tag{11}$$

By setting the threshold value, abandon the point when $\left|D(\hat{X})\right|$ is less than the extreme value point of threshold value. A typical threshold size can be set to 0.05.

3.3 OUT PART OF THE EDGE POINTS

In order to obtain stable characteristics, just out the poor contrast with extreme value point is not enough. Usually, this kind of edge pixels in is near to have larger curvature on the edge of the direction, but on the edge of the vertical direction is opposite. The curvature of key point can be calculated by a Hessian matrix H with 2×2:

$$H=\begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}. \tag{12}$$

The expansion term in equation is calculated by difference value of adjacent sample point.

It has proportional relationship between the eigenvalues of Hessian matrix H and the principal curvatures of D. Set α and β as the maximum eigenvalue and minimum eigenvalue of matrix H respectively, then we can get the trace $Tr(H)$ and determinant $Det(H)$ of H:

$$Tr(H)=D_{xx}+D_{yy}=\alpha+\beta, \tag{13}$$

$$Det(H)=D_{xx}D_{yy}+D_{xy}^2=\alpha\beta. \tag{14}$$

If $Det(H)$ is minus, then out the key point. Set γ as the ratio of the maximum and minimum eigenvalue, namely $\alpha=\gamma\beta$, we can get the following equation:

$$\frac{Tr(H)^2}{Det(H)}=\frac{(\alpha+\beta)^2}{\alpha\beta}=\frac{(\gamma\beta+\beta)^2}{\gamma\beta^2}=\frac{(\gamma+1)^2}{\gamma}. \tag{15}$$

The results of the equation only depend on the ratio of the eigenvalues γ, and have nothing to do with two eigenvalues. When two eigenvalues equality, namely $\gamma=1$, $\frac{(\gamma+1)^2}{\gamma}$ achieves the minimum, $\frac{(\gamma+1)^2}{\gamma}$ with the increase of γ.

Seen from the above analysis, relying on the extracted features in great quantities, SIFT algorithm with information redundancy strategy to achieve precise matching, and at the same time also increases the amount of calculation of the algorithm itself. For the tennis video feature extraction, it often requires system can real-time processing the video, so we improve it.

3 The improved SIFT algorithm based on tennis video feature extraction

3.1 OPTIMIZING THE DISTRIBUTION OF FEATURE POINTS BASED ON TEXTURE FEATURES

In order to make the feature points can be evenly distributed in each set of different scales of video image (N_{ol}), so as to improve the scale-invariant features of SIFT algorithm, then it need to allocate the scale video image feature points with a certain proportion for each group of each layer. We limit the number of feature points by adding the image texture feature. Set the total number of predefined needed feature points as N, then the number of feature points allocated by the corresponding 0 group of the first layer l scale images is N_{ol}, and define the corresponding obtained proportionality coefficient as F_{ol}:

$$N_{ol}=N\cdot F_{ol}, \tag{16}$$

$$\sum_{o=0}^{ON-1}\sum_{l=0}^{LN-1}F_{cl}=1. \tag{17}$$

As the video images in the scale space after Gaussian function of smoothing, the number of feature points will reduce with the increase of scale. So, proportionality coefficient F_{ol} is inversely proportional to with the scale coefficient. If $f_0=F_{00}$, means the proportion to the control points in 0 group 0 layer of the scale video image, the corresponding scale coefficient is SC_{00}, then each group of each of the video image feature points percentage can be expressed as follows:

$$F_{ol} = \frac{SC_{00}}{SC_{ol}} f_0, \tag{18}$$

$$F_{ol} = \frac{f_0}{k^{LN \cdot o + l}}, \tag{19}$$

where $o = 0,1,2,...,ON-1; l = 0,1,2,...,LN-1; k = 2^{1/LN}$.

Finally using Equations (16) and (17):

$$\sum_{o=0}^{ON-1} \sum_{l=0}^{LN-1} \frac{f_0}{K^{LN \cdot o + l}} = 1 \Rightarrow f_0 = \frac{k^{ON \cdot LN - 1}}{\sum_{n=0}^{ON \cdot LN - 1} k^n}. \tag{20}$$

In addition to the uniform distribution on the scale space, the same demands on the arrangement of video images were distributed evenly, so based on video image area was divided into small units n_cell , each unit allocates with corresponding proportion of feature points, to achieve uniform distribution of feature points on the image space.

However, such a set makes use of the same group of class number, the number of feature points in video image sub-block preset is fixed, that causes the characteristic quantity which extract by the small pictures not very obvious texture feature greatly reduced, and affects the accuracy of registration, moreover the too much characteristic quantity acquired by complex texture feature of images characteristics and then affects the speed of whole system.

Video image texture complexity can be described by information entropy of video image, so first of all statistics for the amount of information of video images, through the statistical results in the characteristic quantity of the default values. The predefined number of feature points is inversely proportional to the video image texture features, and is proportional to the size of the image:

$$1500 < CP_{num} = \frac{image\_size}{image\_entropy} \times 5\% < 4000. \tag{21}$$

3.2 THE OPTIMIZING FEATURE EXTRACTION AND MATCHING BASED ON PROJECTION TRANSFORMATION

Feature points matching uses two steps to complete, first to use the minimum Euclidean distance initial matching, then to remove the "point" of the feature point matching by using projection transformation model.

3.2.1 Initial matching of feature points

Feature points of similarity measure usually adopt the method of distance measurement, such as the Euclidean distance, Markov distance, etc. The Euclidean distance as a method to measure, first using the method achieves two points p_1 and p_2 near the feature point p , calculates Euclidean distance $Ed(p, p_1)$ and $Ed(p, p_2)$ of p . If the ratio of $Ed(p, p_1)$ and $Ed(p, p_2)$ is less than the threshold T_d , then p and p_1 as a pair of matching feature points:

$$\frac{Ed(p, p_1)}{Ed(p, p_2)} < T_d. \tag{22}$$

2.2.2 Out the "point"

On the premise of ignoring imaging abnormalities, the video image of different perspectives in the same scene has a one-to-one relationship. In homogeneous coordinate, the image $X(x, y, 1)^T$ and $X'(x', y', 1)^T$ meet the relationship of projection transformation:

$$X' \sim HX = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix} X. \tag{23}$$

where "~" sets as proportional relationship with left and right. matrix H has 8 independent variables, the projection transformation relations for specific equation as follows:

$$x' = \frac{h_0 x + h_1 y + h_2}{h_6 x + h_7 y + h_8}, \tag{24}$$

$$y' = \frac{h_3 x + h_4 y + h_5}{h_6 x + h_7 y + h_8}. \tag{25}$$

From the equations above only four control points to corresponding matching are required to be received to calculate the space transform relation between images. And then it is possible to obtain feature points of maximum satisfy geometrical model by using random sampling consistency algorithm to iterative operation. Transformation model between images is calculated by using the least square method:

$$Ax = b \Rightarrow x = [A^T A]^{-1}. \tag{26}$$

To sum up, the concrete steps to the optimizing feature point extraction and matching based on projection transformation as follows.

1) First, determine the total number N of the need feature points: predefined extra feature points can reduce the operational efficiency of the algorithm, and too few feature points can affect the accuracy of registration. According to the ratio of 0.6% ratio of the input image size and the original image information entropy to determine the total number of feature points. But with the increase of the input image, the total number of feature points requires less than 4000, and picture too small the total number of feature points requires more than 2000, in order to ensure the speed of operation and the accuracy of registration;

2) Building a DOG pyramid: according to the predefined number of feature points (N_{ol}) extract space extremum points from each floor in each group of the

image, the images divided into n_cell areas, the number of feature points for each area as follows:

$$n\_cell_i = 3\frac{N_{ol}}{n\_cell}. \quad (27)$$

3) Filtering principal curvature of the feature points in threshold $T_r = 10$;

4) Calculating the information entropy of feature points

and its radius is 3σ, ordering it from big to small, then keeping the former n_cell_i feature points;

5) Describing characteristics of feature points which extracted by the standard SIFT algorithm;

6) Matching feature points by using the minimum Euclidean distance and projection model.

In order to avoid the reduction of number of control points, other block or scale image is extracted to compensate where the number of control points cannot reach a predefined.

After completing the extraction of feature on tennis video image, use the minimum RMSE to determine the accuracy of registration:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}((x_i - \bar{x}_i)^2 + (y_i - \bar{y}_i)^2)}{N}}, \quad (28)$$

where (x_i, y_i) sets as the coordinates of the reference image, $(\bar{x}_i, \bar{y}_i)$ as the coordinates after the input image projection model transformation, Obviously the smaller the value represents the higher registration precision.

4 The performance simulation of algorithm

In order to verify the effectiveness of the proposed improved algorithm, simulation experiments on it. First of all, testing the improved SIFT algorithm performance, the experimental environment is: The CPU is Pentium(R) Dual-Core CPU E5300 2.60GHz, internal storage is 2GB, video card is NVIDIA GeForce 9300GE. Extracting and matching the feature of video images with different resolution. The results show as follows:

TABLE 1 The results of feature extraction based on different pixel image

Image pixels	Feature points extraction	
	SIFT	IM-SIFT
100×100	24	35
200×200	46	73
300×300	61	103
400×400	84	112
500×500	94	141
600×600	101	169
700×700	115	188
800×800	130	203
900×900	137	214
1000×1000	142	246

TABLE 2 The matching results of characteristics based on different pixel image

Image pixels	The exact value		The matching time	
	SIFT	IM-SIFT	SIFT	IM-SIFT
100×100	64.3%	96.3%	0.34	0.13
200×200	67.2%	91.2%	1.25	1.03
300×300	61.4%	96.7%	2.58	1.57
400×400	59.3%	98.0%	5.26	3.14
500×500	56.2%	99.1%	9.81	5.14
600×600	64.6%	98.5%	15.85	8.37
700×700	58.6%	97.3%	30.42	11.03
800×800	58.8%	97.6%	71.35	14.13
900×900	61.6%	93.2%	140.45	21.46
1000×1000	66.7%	91.5%	190.91	25.16

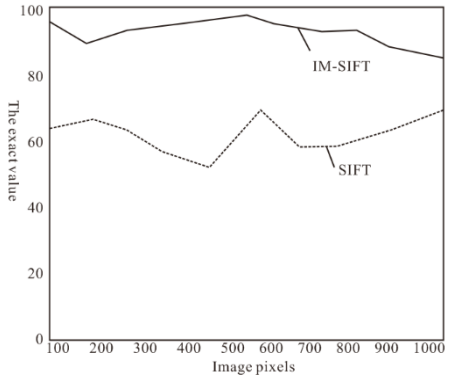

FIGURE 1 The accuracy comparison of image feature extraction

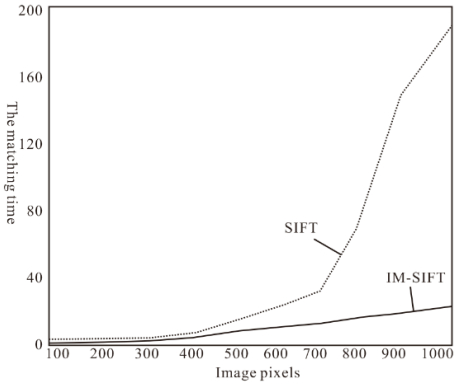

FIGURE 2 Time comparison of image feature extraction

Then the improved SIFT algorithm is applied to feature extraction and matching in tennis video, the matching error statistics are as follows.

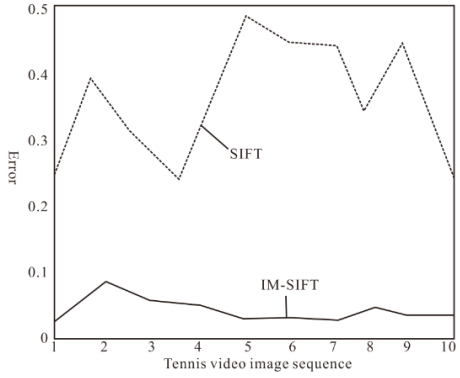

FIGURE 3 The feature matching error comparison in tennis video

Seen from the above simulation results, the improved SIFT algorithm for video image feature extraction and matching, with a faster speed and accuracy, can be very good application in Wang Qiang video feature extraction and matching.

5 Conclusions

Sports video data is a kind of important media data, it has a wider audience. People not only satisfy the direct viewing and simple browsing sports video but with more diversified needs, such as abstract of wonderful snippets, a detection of specific event, etc. The improved SIFT algorithm based on tennis video feature extraction and matching is proposed in this paper, seen from the simulation results, it has faster speed, and less error in the tennis in the video feature extraction and matching.

References

[1] Zhang X, Zhi M 2013 Adaptive threshold based slow motion replay detection method for tennis video *Computer Engineering & Science* **35**(4) 99-103

[2] Wang P 2013 Tennis Video Analysis Using Transformed Motion Vector Field *Acta Electronica Sinica* **33**(5) 935-8 *(in Chinese)*

[3] Wu P 2014 Sports Videos Classification Based on Color Texture and SVM *Journal of Fujian Teachers University (Natural Science)* **30**(2) 34-41

[4] Zhu Y 2013 Sports Video Classification Based on Marked Genre Shots and Bag of Words Model *Journal of Computer-Aided Design & Computer Graphics* **25**(9) 1375-83

[5] Zhang H, Yang X, Huang C 2012 Tagging And Indexing Sport Video Based on Hierarchical Semantics *Computer Applications and Software* **29**(10) 258-60

[6] Song G 2014 Video Classification Based on Region Features and HMM *Journal of Southwest China Normal University (Natural Science)* **35**(2) 180-4

[7] Zhu Y 2013 Analysis and Detection of Redundancy Data on Basketball Video *Mini-micro Systems* (9) 1837-78

[8] Hou L 2014 Method for Slow-motion Replay Detection on Compressed Domain in Sports Video *Computer Science* **36**(9) 283-6

[9] Lao Songyang, Bai Liang, Liu Haitao, Alan F Smeaton. (2014) Semantic Content Analysis Model for Sports Video Based on Perception Concepts and Finite State Machines. *Mini-micro Systems*, (6), 1137-1141.

[10] Xu H, Xiao H, Hou H 2014 Algorithm for Maneuvering Target Tracking in Sports Video Frequency Based on IMM *Journal of Image and Graphics* **14**(5) 920-4

[11] Ji S 2014 System Identification Based Sports Video Recognition *Journal of Wuhan University of Technology* (14) 142-4

[12] Bu Q, Hu A 2014 An Approach to User-Orientated Highlights a Sport Video *Pattern Recognition and Artificial Intelligence* **21**(6) 782-6

[13] Han B 2014 Enhanced Sports Video Shot Transition Detection Based on a Unified Feature Model *Video Engineering* **31**(8) 76-9

[14] Liu Y 2014 Playfield Detection Using Adaptive GMM and Its Application in Sports Video Analysis *Journal of Computer Research and Development* **43**(7) 1207-15

[15] Zhang L 2014 Support Vector Machine (SVM) Meta Classifier Based Sport Video Classification *Journal of Beijing Institute of Technology (Natural Science Edition)* **26**(1) 41-4

Authors

Yang Wang, June 1978, China.

Current position, grades: Langfang Normal University Sports Institute.
University studies: master's degree at Beijing Sports University in 2010.
Scientific interests: sports teaching and training.
Publications: 6.

Haiyan Geng, 20.06. 1977, China.

Current position, grades: Langfang Normal University Sports Institute.
University studies: master's degree in education at Hebei Normal University Sports Institute in 2011.
Scientific interests: sports teaching and training.

TianLin Geng, February 1985, China.

Current position, grades: Baoding Gao Yang Hongrun Middle School.
University studies: bachelor's degree at Hebei Institute of Physical Education in 2005.
Scientific interests: sports teaching and training.
Publications: 2.