Journal of Network Intelligence ©2024 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **9**, Number **4**, November **2024**

Player Identification Method Based on Machine Learning in Video Sequence

Yong-Xiao Li

The Sports Academy
Jiangsu Ocean University, Lianyungang 222005, P. R. China
862193331@qq.com

Ke Zhao*

Graduate School
Harbin Sport University, Harbin 150006, P. R. China
The Sports Academy
Zhangjiakou University, Zhangjiakou 075000, P. R. China
k435934108@outlook.com

Cai Feng

College of Management Science
University of Cyberjaya, Persiaran Multimedia, 63000 Cyberjaya, Selangor, Malaysia
nt2502@163.com

*Corresponding author: Ke Zhao

Received December 15, 2023, revised March 11, 2024, accepted June 22, 2024.

Abstract. *By identifying a player's number, it is easier to collect and count a player's personal data, such as playing time, number of passes, number of shots, points, fouls, etc. These data can be used to evaluate players' performance, identify players' strengths and weaknesses, and provide suggestions for training. However, with the increase in the number of sports videos, the traditional video target recognition methods based on manual annotation can no longer meet the requirements of work efficiency. Therefore, this work proposes a method using computer vision and machine learning to detect and recognise players' numbers from sports video sequences. First, the sample images from the input video sequences are preprocessed, including Gaussian blurring and greyscaling. Then, a jersey region localisation based on Directional gradient Histogram (HOG) features and SVM is proposed in order to quickly detect key regions containing names and numbers. Secondly, the Shuffle module is introduced in LeNet-5 to improve the performance of the model. By inserting the Shuffle module between the convolution and pooling layers of LeNet-5, the interaction between feature maps can be increased. Finally, a basketball game video of the 2013 NBA Lakers vs. Bucks was used for testing. The results show that the jersey region localisation method based on HOG features and SVM is robust to background interference and noise, and the accuracy of localisation reaches 91.2%. Compared with BPNN, LeNet-5 and LSTM, the improved LeNet-5 has the highest average accuracy of 95.5% under complex conditions.*

Keywords: Machine learning; Sports video; Number recognition; HOG; SVM; LeNet-5

1. **Introduction.** With the explosive growth of video data, traditional manual annotation-based video analytics can no longer meet the demand of massive data retrieval due to its many limitations, and people have started to focus on content-based video analytics [1, 2, 3]. Video is a fusion of images, text, sound and other multimodal media data,

the data has an unstructured organization, the image is stored in the form of pixels to represent the object's colour, brightness and other underlying semantic information, the amount of information, the performance of a variety of content, but the lack of high-level semantics of the video understanding, so how to achieve the computer to automatically extract the high-level semantic content from the video data of the research has become a video retrieval, hotspot in the fields of intelligence and automation [3, 4].

Content-based video analytics, which has great commercial value in the field of sports videos, exploring analytics for sports videos has also become a hot research topic in recent years [5, 6]. At the same time, the traditional one-to-many broadcasting model cannot satisfy the different needs of viewers, for example, a James Brown fan may want to watch the whole game, while another fan just wants to see James Brown's dunk and cap moves only, and the analysis of sports videos must satisfy the user's personalised needs. Among them, the detection and automatic identification of players in sports videos play an important role in the content analysis, retrieval and personalised recommendation of sports videos, which is the basis of intelligent sports video analysis [7, 8] and has high practical significance.

By identifying the numbers of players in sports videos, the tactical patterns of teams can be observed and analysed more clearly [9], such as formations, attack routes, defensive zones, and coordination methods. These data can be used to develop and optimise game strategies, identify the weaknesses of opponents and their strengths, and improve the game winning rate. Player identification methods can enrich and beautify the live broadcast and replay effects of the game more, such as adding number tags to players, displaying players' names and data on the screen, and zooming in on players at important moments [10]. These effects can improve the spectacle and interest of the game and attract more viewers and fans. Player identification methods can also be used to referee and manage matches more accurately and fairly, e.g., in case of disputes, the numbers of the players involved can be quickly replayed and confirmed to avoid misjudgments and omissions and to maintain the fairness and order of the match.

In summary, how to effectively achieve the positioning and identification of players has very important research value and practical significance. However, due to the more intense physical confrontation in the game, the player's body twisting amplitude is very large, and the jersey number has a drastic deformation, the positioning and identification of the player is difficult. The existing methods do not detect the position of players quickly and the recognition rate is not high, while there are still some difficulties in player detection and recognition due to problems such as being affected by light and the audience being too close to the pitch [11, 12]. Therefore, seeking methods for automatic player recognition in sports videos is still a difficult and a major challenging topic in sports video analysis at present.

1.1. **Related Work.** players in sports videos are recognised by two main methods: face recognition and subtitle recognition.

Currently the most researches are based on face recognition techniques, a common method is achieved by detecting and recognising faces in video frames. Zhang et al. [13] proposed a deep learning based face recognition method for player recognition in sports videos. The method utilises face detection and alignment along with multi-task learning and attention mechanisms to improve the accuracy and robustness of recognition. The method utilises a deep convolutional neural network to process close-up shots of the player's face, which can achieve high recognition accuracy but is limited to specific shots. Zhou and Zhang [14] proposed a face recognition method based on image segmentation and feature fusion for player recognition in sports videos. The method uses image segmentation to separate face regions from complex backgrounds, and then achieves automatic scene classification as well as player annotation by combining face recognition with a video grammar model, which improves the efficiency and stability of recognition. However, this method of combining contextual information relies on complex syntactic modelling although it can improve robustness. Kong et al. [15] proposed a face recognition method based on face keypoints and spatio-temporal features for athlete recognition in sports videos. The method uses face keypoints to locate the pose and expression of the face, and then uses spatio-temporal features to capture the dynamic changes of the face, which improves the sensitivity and discrimination of recognition. This is a bimodal approach, using both facial features and jersey numbers for decision making. However, while this multi-source information fusion recognition framework improves the fault tolerance of the system, the actual deployment requires simultaneous acquisition of multiple types of information in the image. However, face-based recognition techniques can only recognise video frames that have a positive face in the close-up shots and there are many limitations in recognising players directly by their faces due to issues such as player posture, field lighting and occlusion in sports.

Player recognition based on subtitle recognition detects and recognises scene text and superimposed text in videos. Lu et al. [16] use recognised subtitle text information for sports video footage classification. However, this method is only applicable to text fixed on the surface of rigid objects, whereas player numbers are on non-rigid surfaces such as jerseys. In addition to recognising players through faces and subtitles etc., recognising player numbers is also a method of identifying players. The jersey number is an important identifier for players in sports, especially in team sports, where the number is mainly on the front or back of the jersey, and sometimes the player's name is attached to the number for the purpose of easy identification by the referee and the other players. Šaric et al. [17] used a player jersey colour model to locate the player and then identify the number to identify the player, which requires human intervention. Messelodi and Modena [18] proposed to use generalised learning vector quantisation to obtain player numbers and KNN clustering to identify them.

1.2. Motivation and contribution.

The faces of the players in the sports video sequences are not clear and it is not possible to identify the players by face recognition, but only by number. Therefore, this work adopts the method of locating the player first and then extracting the jersey number. Since sports video sequences contain full-body images of multiple players, this paper utilises Directional gradient Histogram (HOG) [19, 20] features, which are commonly used for human detection, to locate players. The jersey region fine localisation is achieved by training SVM classifiers [21, 22]. Finally, the final identification of numbers is performed in the localised jersey regions. The main innovations and contributions of this work include:

(1) Aiming at the problem that the traditional region localisation algorithm has poor localisation effect when dealing with more interference information or poor quality of sample images, a combined localisation algorithm based on colour information, edge detection measure, HOG and SVM phase.

(2) Introducing Shuffle modules in LeNet-5 [23, 24] to improve the performance of the model. The interaction between feature maps can be increased by inserting Shuffle modules [25] between the convolution and pooling layers of LeNet-5. In addition, by adjusting the structure of the module to maximise the retention of important information in the image, the new module is called R-Shuffle.

2. Pre-processing of sample images.

2.1. Gaussian blurring.

Gaussian blurring is a method of blurring an image by taking a weighted average of the neighbouring pixels for each pixel in the sample image, which is mainly used in the pre-processing stage of an image to remove noise from the image, especially before extracting edge features to remove irrelevant details from the sample.

Gaussian fuzzy algorithm is similar to the simpler average fuzzy algorithm, but because the image is continuous if the use of simple summation of the average will make the sample image to lose a lot of important information, so the Gaussian fuzzy algorithm on the basis of the use of the weighted average of the closer to the target pixel points and closer to the target of the point of the weight is set to the larger, and the further away from the target pixel point of the weight of the point of the weight is set to the smaller and the distribution of the weights of these surrounding pixels is equated with the values of a two-dimensional Gaussian distribution, while the assignment of weights is achieved using a two-dimensional normal distribution, as shown in Figure 1.

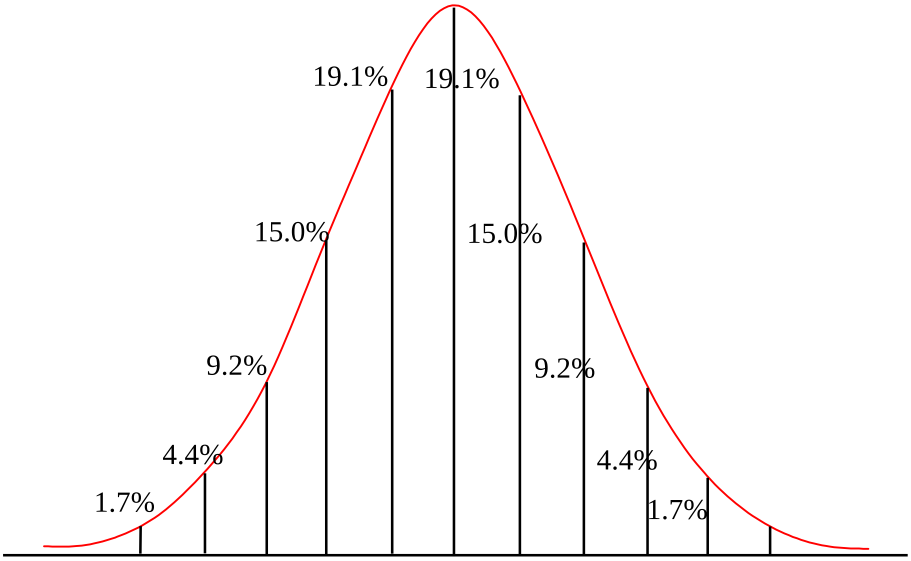


Figure 1. Normal distribution

The density function of a two-dimensional normal distribution is the Gaussian function we need, and its one-dimensional form is shown below:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

where u denotes the mean and σ denotes the standard deviation.

In calculating the mean, the center point is the origin, then $u = 0$. Therefore, Equation (1) can be simplified as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \tag{2}$$

The two-dimensional Gaussian function can then be derived from the simplified one-dimensional Gaussian function as:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

Finally, by setting the. The value of the weight matrix is obtained, and then the value of the Gaussian blur is calculated from the weight matrix, and the image after Gaussian blurring can be obtained by repeating the above method continuously.

In this paper, Gaussian filtering based on colour images is performed before edge feature extraction, which can remove noise for the next edge feature extraction, while retaining the colour information of the samples as much as possible, which will make it easier to find the edge points than using greyscaling directly. After experimentation, it was found that the rest of the filters such as median and linear filters were not very helpful for edge feature extraction, so these two methods were not used. The setting of Gaussian fuzzy radius also has a very big influence on the extraction of edge features, too big and too small will cause the number can not be located, after many tests, the final choice of Gaussian radius of $R = 4$ when the jersey region has the best localisation effect.

2.2. **Grayscaling of sample images.**

Currently, the captured samples obtained through cameras in sports video sequences are colour images based on the RGB model. In the colour images under this model, each pixel point consists of red (R), green (G) and blue (B) components superimposed in different ratios, and each primary colour component can take a total of 256 grey scale variation values from 0 to 255, which provides a sufficient number of extractable features for searching and determining the region containing the player's identity information.

However, this is obviously too much data for the use of edge feature extraction to determine the player number, whereas after greyscaling the sample image it still maintains the key features of overall or local chromaticity and luminance, and the use of this feature is sufficient to determine part of the jersey identity region. Therefore, after the Gaussian filtering process, the colour image is usually not processed directly, but is greyscaled and then the edge features are extracted, thus completing the coarse positioning of the jersey identity region.

The greyscaling of an image is the unification of the three primary colour components under the RGB model, and there are several commonly used methods as follows:

(1) Component method. The three primary colour components of each pixel point are selected according to different needs, and one of them is used as the grey value of this pixel point in the greyscale image.

$$Gray_1(i,j) = R(i,j), \quad Gray_2(i,j) = G(i,j), \quad Gray_3(i,j) = B(i,j) \tag{4}$$

(2) Maximum value method. The largest of the three primary colour components of each pixel point is taken as the grey value of this pixel point in the greyscale image.

$$Gray(i,j) = \max\{R(i,j), G(i,j), B(i,j)\} \tag{5}$$

(3) Weighted average method. Since the human eye is sensitive to colours in the order of green, red, and blue, the weighting coefficients for weighted averaging were empirically taken as $a = 0.299$, $b = 0.578$, and $c = 0.114$.

$$Gray(i,j) = a \cdot R(i,j) + b \cdot G(i,j) + c \cdot B(i,j) \tag{6}$$

Obviously, both the first and second methods cause a large amount of key information to be lost, making subsequent feature extraction difficult. Therefore, the third method (weighted average) is chosen in this work, which is able to retain the desired key features to the maximum extent through Gaussian filter denoising and greyscaling, and is empirically proven to have good results.

3. **Jersey region localisation based on HOG features and SVM.**

3.1. **Introduction to the principle.** Whenever there is a great goal or a major event in a game, a close-up shot is often given to the relevant player, and the identification of the player in these close-up shots plays an important role in the analysis and retrieval of the video. In these close-ups, the player is almost always in the centre position, often containing only the upper body of the player, and the overall detection method cannot be used to locate the player. Due to the close proximity of the player to the camera, the jersey number and the team name above provide a wealth of textual information. However, to accurately identify the jersey number and the team name above it, the jersey area must first be accurately located.

The jersey region location localisation method adopted in this paper is not a traditional algorithm based on single feature extraction, but a multi-feature localisation method based on threshold judgement. This method is suitable for the case where the sample image works well, and at the same time, it can maintain a high recognition efficiency under the interference of various objective conditions. Firstly, the region similar to the base colour of the jersey is extracted from the sample image, and then edge information is detected if there are too many interfering factors or too few candidate regions by threshold judgement, and then the candidate connected domains are formed by using morphology and custom region linking, and then the candidate regions are filtered according to the a priori information of the jersey size in order to complete the initial localisation of the jersey region.

3.2. **The HSI model.** Both the HSI model [26] and the RGB model are able to describe colour images, and there is a corresponding mathematical transformation between them that can be converted to each other, which is in fact a conversion from a unit cube based on a right-angled coordinate system to a bipyramid based on cylindrical polar coordinates. However, the RGB model can only describe the numerical values of colours, but not the colours interpreted by humans. Therefore, the HSI model imitates the way of human recognition of coloured objects, and describes the colour image by the colour space composed of the three parameters of Hue, Saturation and Intensity, and can show the actual change of these three parameters through a biconical space model, whose biconical space model is shown in Figure 2.

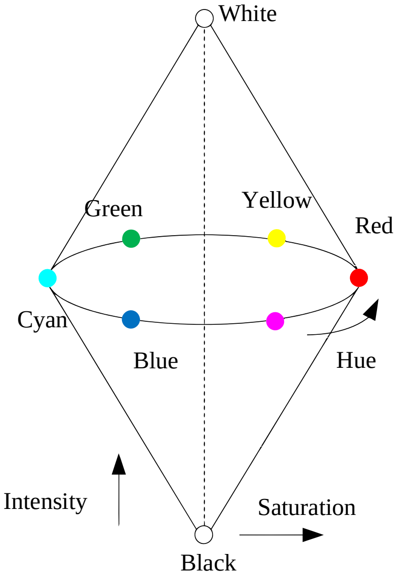


Figure 2. Biconical space of the HSI model

The acquired sample images under the RGB model are transformed into a three-parameter description mode based on the HSI model through mathematical relational equations. Then, according to the actual situation of the jersey, the parameter setting template applicable to the jersey is determined through actual testing, which achieves a good recognition effect and has good general applicability. The calculation process of conversion from RGB model to HSI colour space is shown as follow:

$$\theta = \arccos\left(\frac{[(R-G)+(R-B)]}{2\sqrt{(R-G)^2+(R-G)(G-B)}}\right) \tag{7}$$

Once the value of θ in Equation (7) has been calculated, the pixel values of the three colour components in the RGB mode can be calculated as the specific values of Hue (H), Colour Saturation (S), and Brightness (I) as shown as follows:

$$H = \begin{cases} \theta & B \le G \\ 360 - \theta & B > G \end{cases} \tag{8}$$

$$S = 1 - \frac{3}{(R+G+B)}\left[\min(R,G,B)\right] \tag{9}$$

$$I = \frac{1}{3}(R+G+B) \tag{10}$$

According to the above four equations, the three parameter values describing the image information under the HSI-based model can be obtained, and the subsequent need to traverse the sample image. When the pixel value of hue (H) is between 195–270 or 29–75, and the values of the two parameters, colour saturation (S) and luminance (I), are between 0.37–1.1, the sample image is calibrated in black and white binary by using the data templates obtained from the experiments.

3.3. **Edge detection.** The theoretical and mathematical basis of image edge detection is more mature, and it is used to bring out the edges of the jersey number region by detecting the parts of the sample image that have a large change in brightness. By extracting these edge features, the processing of the whole sample image can be avoided and the processing speed of the player identity number location determination can be accelerated. However, in the case of the jersey number region near the presence of more lines of information, only edge feature detection will result in a significant reduction in the accuracy of the positioning, so it is necessary to carry out multi-feature extraction in order to ensure that the positioning accuracy, and the following part of the edge detection methods will be Some of the edge detection methods will be introduced in the following.

(1) Canny edge detection.

Canny edge detection is a more effective feature extraction method, which is usually divided into four steps: eliminating noise, calculating gradient magnitude and direction, non-maximum value suppression and hysteresis threshold, thus creating the problem of implementing the algorithm in a more complex process. Before extracting the edge features, it is necessary to use a filter for convolutional noise reduction as follows.

$$G(x,y) = \frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{11}$$

Then a pair of convolution arrays G_x and G_y, acting in the x and y directions, respectively, are set up first to compute the horizontal and vertical differentials, and then the gradient magnitude and direction are computed using Equation (12).

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{12}$$

$$G = \sqrt{G_x^2 + G_y^2}, \theta = \arctan\left(\frac{G_y}{G_x}\right) \tag{13}$$

Next, the gradient strength of the current point is compared with the positive and negative gradient direction points one by one, and the point with the largest gradient strength in the same direction is obtained and retained, while the rest of the points are suppressed. However, after this process the sample image will still be dark due to the many small gradient modulus points, so two lag thresholds (high and low) need to be used. Pixel points higher than the high threshold need to be retained, while pixels below the low threshold need to be excluded, and then pixel points that lie between the two thresholds and are only connected to points above the high threshold need to be retained.

(2) Sobel operator.

To extract edge features using the Sobel operator, it is necessary to first convolve the image to be processed I with an odd-sized kernel in the x and y directions, respectively, as shown below:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \cdot I, G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \cdot I \tag{14}$$

The above two results are then combined to find the approximate gradient for each point in the image. The direction of the gradient can be solved inversely by using the cotangent values of the vertical and horizontal directions.

However, Sobel operator also has some problems, for example, in the case of the sample image with more abundant line information, it will not be able to accurately detect the jersey number area, so it is necessary to extract a variety of features in the localisation process, in order to achieve the precise location of the identity number.

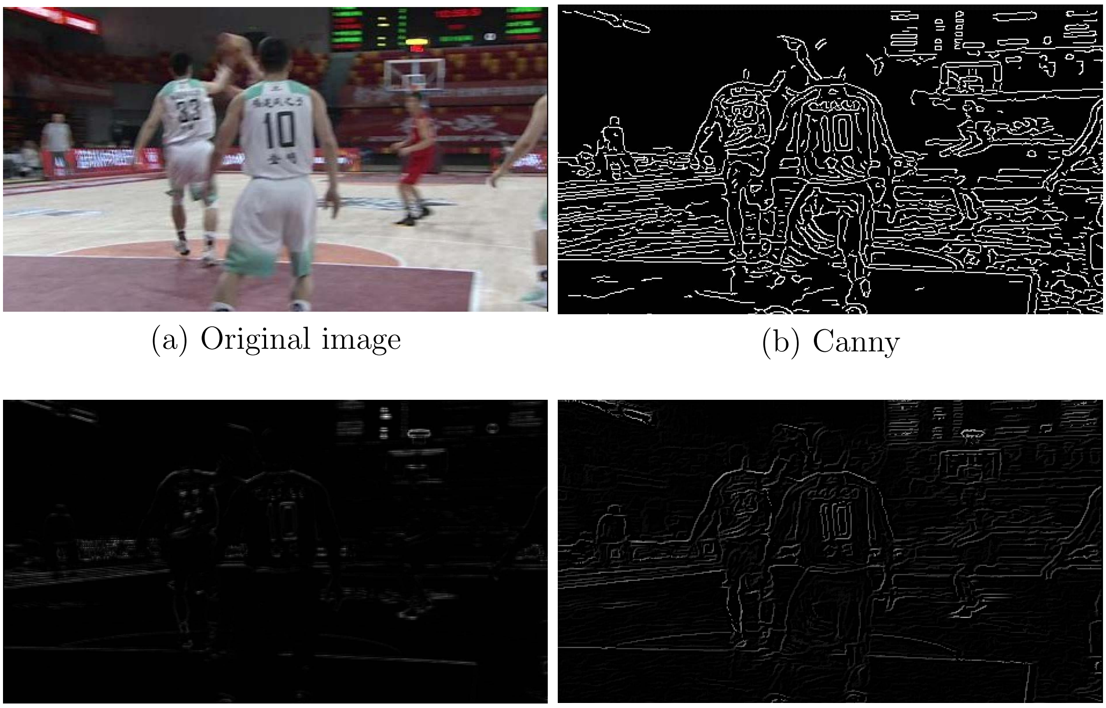
(3) Laplacian operator.

The detection of edge information using Sobel operator is based on the fact that the pixel value in the edge part jumps, so the derivatives in the edge part will be extreme. The Laplacian operator is based on the fact that when the first-order derivatives are extreme in the edge part, the second-order derivatives will be zero to obtain the edge information. From the above principle, we can know that the second-order derivatives can be used to detect the edge information of the image, and by using the Laplacian operator to derive the derivatives of the two directions of the two-dimensional image, the process of calculation can be made easier. The definition of the Laplacian operator is shown as follows:

$$Laplace(f) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{15}$$

In practical applications, the Laplacian operator obtains the Laplace transform result by using the derivatives in the horizontal and vertical directions calculated by the Sobel operator. It can highlight some small detail information while preserving the background of the image, which is beneficial to the extraction of edge features.

The detection results of the three edge detection operators are shown in Figure 3, and considering the processing effect, processing speed and recognition efficiency, the Laplacian operator is finally adopted for edge feature extraction. In this paper, after the threshold judgement in the jersey number coarse positioning algorithm, the original RGB sample image is first Gaussian filtered to remove as much noise as possible. Then the Laplacian edge feature detection, and here there is no first-order derivation, but the use of the peripheral value of the weighted average of the method of convolution, so as to smooth the processing of the noise of the image, and more accurately retain the identity number of the edge information.



(a) Original image (b) Canny

(c) Sobel (d) Laplacian

Figure 3. Test results of different edge detection algorithms

3.4. **Mathematical Morphology Operation.** After the above series of processing, the color and edge interference information not related to the jersey number region is basically eliminated, but the edge information of the number region is more sparse, so the binary image needs to be processed by closing operation and custom region connection.

In this paper, we use two morphological operations, erosion and expansion, to transform the black pixels of the numbered regions, and then use a custom region joining method to form a connected block. The advantage of this is that the two morphological operations can reduce the number of traversals of the custom region joining, and the morphological operations can also convert the gaps of black pixels that cannot be filled by the custom region joining method to white pixels, forming a connected domain for easy extraction.

(1) Corrosion. The erosion operation involves placing the center of an $m \times n$ rectangular region over each pixel point of the image and replacing the value of this pixel with the minimum value of the pixel point of the surrounding region contained by the template. If the sample image a is eroded by a rectangular box β, denoted as $a \ominus \beta$, its definition is shown as follows:

$$a \ominus \beta = \{x | (\beta + x) \subseteq a\} \tag{16}$$

(2) Expansion. The expansion operation does the opposite, extending and connecting salient points at the edges of the image outwards. It still traverses each pixel point in the sample image using a rectangular template, with the difference that the maximum

pixel value of the surrounding region is used for the substitution operation. If the sample image a is inflated by an $m \times n$ rectangular region β for the expansion operation, denoted as $a \oplus \beta$, it is defined as follows:

$$a \oplus \beta = \{x | (-\beta + x) \cap a \neq \emptyset\} \tag{17}$$

The height and width of the $m \times n$ rectangular region need to be chosen reasonably to improve the result. After a lot of experiments, this paper takes the rectangular region of $m = 17$, $n = 3$ to perform the morphological closed operation on the binary image and achieved good results. However, even with this treatment, there will still be some dots that are not connected or even broken numbered regions, so the next step is to customize the region connection method to process the image after the morphological operation, in order to form a more complete numbered region connectivity domain.

3.5. **Correction of tilted numbers.** Since the context of the study is under complex conditions, it is likely that the outer rectangles of the tilted numbers (pseudo-numbers) will be extracted, but the authenticity number identification is required to be performed under a specific size specification. So in this paper, all the extracted outer rectangles are first subjected to the tilting correction of multiple segments, then to the authenticity number identification, and finally to the processing of the authentic number region.

In order to determine the tilt angle of the numbered regions using the straight line detection method, it is necessary to binarise the region or extract the edge information. However, in the case of complex conditions in the research background, when performing the binarisation process, in order to retain the edge information as much as possible in response to high image brightness and high or low grey values in the non-numbered region, the original image of the location of the region obtained by extracting the customised concatenation domain can first be differentiated. Then, the difference image of the original image is obtained before binarising the region, where the horizontal difference is calculated by the formula:

$$\text{difference}(x, y) = f(x, y + 1) - f(x, y) \tag{18}$$

After obtaining the differential image, the differential image is binarised. The main idea is to set the threshold value T to transform the original image into a non-black or white image to separate the interference from the target, i.e., to maximise the sum of the variances between the two.

3.6. **HOG features.** Histogram of Orientation Gradient (HOG feature) is the statistics of the gradient feature of the local region in the image to be recognised. The gradient represents the direction in which the pixel values change the fastest in an image, and the contours that highlight objects in an image are labelled as edges, so the edges and the direction of the gradient maintain a mutually perpendicular relationship. The model can indirectly recognise objects by the edge contour information, so it can achieve the extraction of edge features and thus indirectly use the contour to recognise objects by obtaining the histogram of the direction gradient in the local region. An example of histogram distribution is shown in Figure 4.

The main idea of HOG feature extraction is to first transform the sample image into a gradient map, and then use a block of $m \times m$ cells (each cell is composed of $n \times n$ pixels) to perform sliding scans of this gradient image in the horizontal (P times) and vertical (Q times) directions at a fixed pixel step size T. Then, the angular range of 2π is divided into D regions within each individual cell, and the cumulative number of gradient values

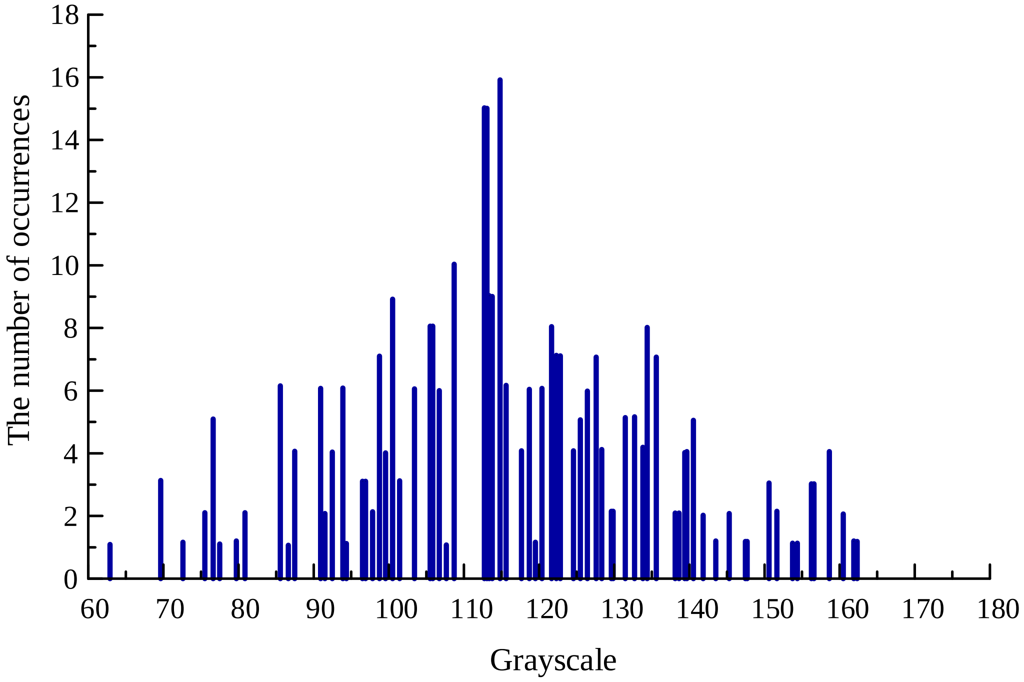


Figure 4. Example of histogram distribution

corresponding to each angular region within this cell is counted to form a D-dimensional feature vector.

After processing as above, a gradient map is transformed into a series of feature vectors that can be processed by the computer (a total of $m \times m \times D \times P \times Q$ features), and according to the relationship between the gradient map and the contour map, the edge contour can be recognised by the computer.

The gradient and gradient direction are computed for each pixel point in the target region in both horizontal and vertical directions, where the gradient operator in the horizontal direction is $[-1, 0, 1]$, while the gradient operator in the vertical direction is $[-1, 0, 1]$, as shown as follows:

$$G_x(x, y) = I(x+1, y) - I(x-1, y) \tag{19}$$

$$G_y(x, y) = I(x, y+1) - I(x, y-1) \tag{20}$$

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \tag{21}$$

$$\theta(x, y) = \arctan\left(\frac{G_y(x, y)}{G_x(x, y)}\right) \tag{22}$$

In the process of obtaining HOG feature vectors of the target region, in order to take into account the processing speed and recognition accuracy, this paper sets the sliding block size to 2×2 and sets the fixed pixel step to $T = 8$. The gradient image is slid-scanned in the horizontal direction for 10 times and in the vertical direction for 2 times, to avoid the situation of unsegmented covered area when the target region is cell segmented, and then obtain the feature description vector of each block. The cell unit is composed of 10×10 pixels and neighbouring cells do not overlap with each other, and then all possible gradient directions are divided into 8 regions ($D = 8$). The features of these blocks are combined to obtain the HOG feature vector of the target region with 640 dimensions ($2 \times 2 \times 10 \times 2 \times 8$).

3.7. **Fine-grained localisation based on SVM classifiers.** SVM is an enhancement of the logistic regression algorithm and understanding logistic regression helps us to understand SVM more clearly.

Logistic regression is one of the methods used in the field of machine learning to solve binary classification problems. Logistic regression assumes that the dependent variable obeys a zero-one distribution, and it can learn a binary classification (0, 1) model based on the features of the input samples. In turn, logistic regression is developed on the linear regression model, which assumes that the linear regression function $H_\theta(x)$ can be expressed as:

$$H_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n = \theta^T x \tag{23}$$

where θ is the training parameter and x is the input data.

After putting the linear regression model into action using the sigmoid function, the hypothesis function $h_\theta(x)$ for logistic regression is obtained expressed as:

$$h_\theta(x) = g\left(\theta^T x\right) = \frac{1}{1 + e^{-\theta^T x}} \tag{24}$$

where g is a sigmoid function.

In this paper, 800 samples containing the jersey number region are selected as positive samples by manual labelling method, and 800 samples such as the background without the number region are selected as negative samples, with a uniformly normalised size of 32×32. Note that in the selection of samples, we try to select the most representative region, and the positive samples should be selected to minimise the interference of the background. The parameters are trained with the help of the machine learning class `CvSVMParams` in OpenCV. The jersey region fine localisation method based on HOG features and SVM is shown in Figure 5.

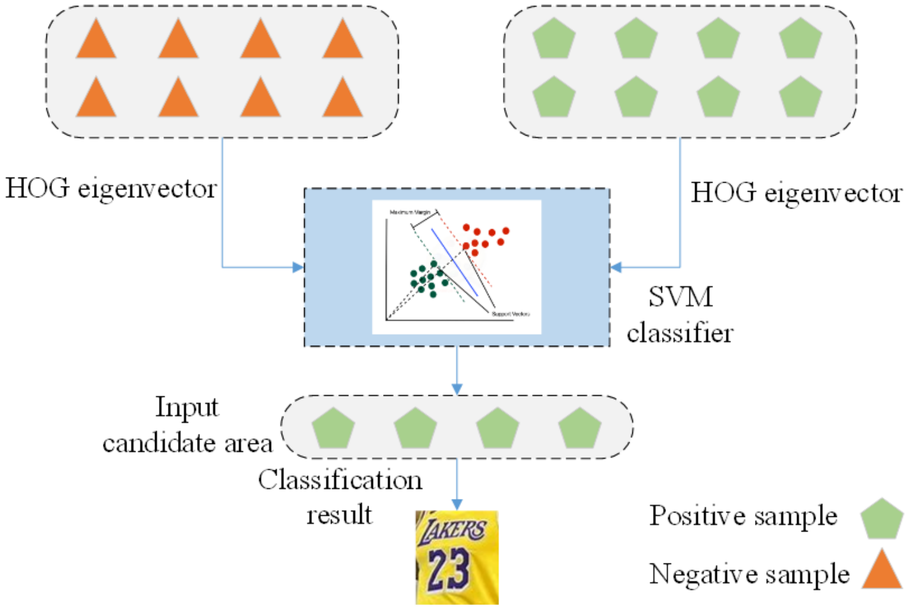


Figure 5. A jersey region fine localisation approach based on HOG features and SVM

SVM models have many parameters to configure, and the selection of parameters is directly related to the prediction accuracy of the final trained model. However, OpenCV provides an auto-training method, the OpenCV SVM library can automatically change the parameters, train and test the model and then select the best ones as the final SVM model parameters.

4. **Player number recognition based on improved LeNet-5.** In sports video sequences, when we locate the player's number region, the number background is the jersey. When the jersey region localisation is satisfied, player number recognition is required. However, due to the different format of the printing of the numbers on each team's jersey, the fonts are often of the team's own design, and the numbers have non-rigid deformation, it is difficult to recognise them directly using OCR software.

For this, we need to build a classifier ourselves. For the recognition of numbers, the commonly used methods are: template matching method, statistical decision making method and neural network. Among them, the template matching method is one of the most commonly used methods, but it is only applicable to print recognition and certain recognition with special restrictions. The statistical decision-making method is highly resistant to interference, but it is particularly dependent on the extraction of features, and the selection of good features is very difficult. Neural networks are a particularly popular method at this stage, especially for deep learning, which does not need to display the extraction of sample features. LeNet-5 is a classical model in the field of deep neural networks [27], which was proposed by Yann LeCun et al. in the early 1990s. LeNet-5 is mainly used for classifying handwritten numeric characters, and is one of the key models in the field of Convolutional Neural Networks (CNNs). one of the key models in the development process.

Although LeNet-5 can perform the number recognition step better in the case of better image quality of the player number region, but in complex conditions it needs to be further investigated to achieve better number recognition efficiency. Therefore, the prototype of LeNet-5 is not directly applicable to player number recognition and needs to be improved appropriately. In this paper, the LeNet-5 model is improved to adapt to player number recognition under complex conditions.

The main structure of the LeNet-5 model consists of a convolutional layer, a pooling layer and a fully connected layer. Through the extraction and dimensionality reduction of the convolutional and pooling layers, LeNet-5 can extract high-level abstract features from low-level pixel features for player number recognition. The input of LeNet-5 is a 32×32 pixel image of the jersey region, and the final output is a 10-dimensional vector.

Shuffle module is a module used to enhance the performance of deep neural networks [28], by rearranging the input feature maps, the nonlinear representation of the network can be improved, thus enhancing the performance of the model. In this paper, we try to introduce Shuffle module in LeNet-5 to enhance the performance of the model. By inserting the Shuffle module between the convolution and pooling layers of LeNet-5, the interaction between feature maps can be increased. However, in some cases, the traditional Shuffle module may lead to information loss or loss of key features [29], as shown in Figure 6. Therefore, in this paper, we design to maximise the retention of important information in the image by adjusting the position of the module, and the new module is called R-Shuffle, as shown in Figure 7. It can be seen that in order to ensure that the Shuffle module retains as much detailed information as possible of the image's jersey, R-Shuffle tries to minimise this loss.

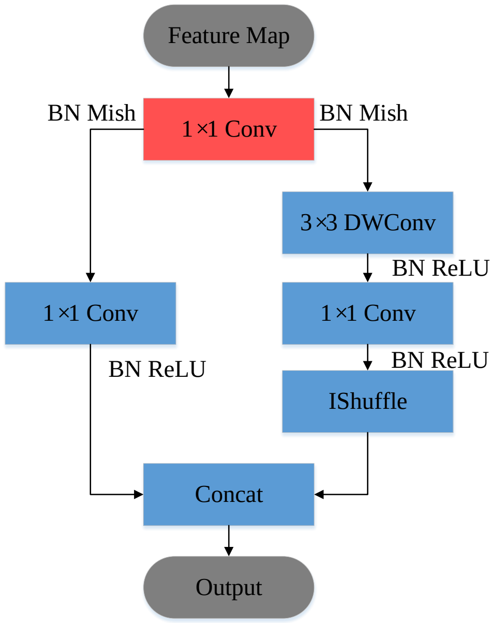


Figure 6. Conventional Shuffle Module

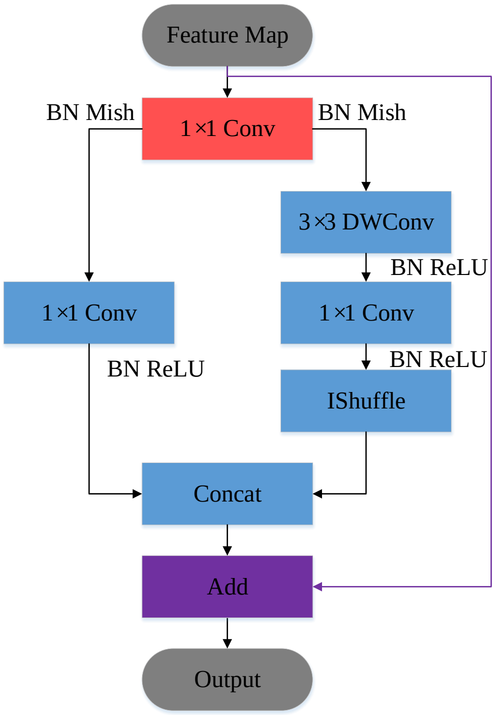


Figure 7. R-Shuffle Module

5. **Experimental results and analyses.**

5.1. **Experimental environment and dataset.** The test environment for data training in this paper is Windows 10 operating system with Intel i7-3720QM @ 2.6 GHz processor and 8GB of internal RAM, the training process took 36 minutes.

Since there is no common database available, the 2013 NBA Lakers vs. Bucks basketball game was used to test this experiment. The frame rate of the video sequence file is 30 frames per second, the resolution of the video is 1280x720 (HD), the video codec format is H.264, the video format is AVI, and the duration of the video file is 20 minutes. The initial colour space of the video file is RGB.

5.2. **Choice of kernel function for SVM..** The kernel function [30] defines the mapping from the input space to the feature space and influences the performance of the SVM on the jersey region fine localisation problem. The authenticity number recognition rates when different kernel functions are chosen are shown in Table 1.

Table 1. Authenticity Number Recognition Rates with Different Kernel Functions Selected

kernel function	**Number of test samples**	**Accuracy %**	**Recall rate %**
RBF	1600	92.67	93.06
Liner	1600	94.21	96.15

It can be seen from Table 1 that Liner is suitable for use when the training samples have a large amount of data and the dimension of each data is also large. although RBF as the mapping function can play the advantage of non-linear model in classification, but when extracting features, if the extracted feature vector in each data dimension is large then its classification effect is not necessarily very good.

5.3. **Jersey region localisation.** The red box indicates the jersey number region determined after passing through the SVM classifier. In the tested video dumps, the accuracy of jersey region localisation based on HOG features and SVM is high as long as the backbone of the player is detected, as shown in Figure 8(c). Figure 8(b) shows the results of Laplacian operator edge detection. The quantitative evaluation results of jersey region localisation are shown in Table 2. It can be seen that the jersey region localisation method based on HOG features and SVM is robust to background interference and noise.

Table 2. Quantitative assessment of jersey area positioning

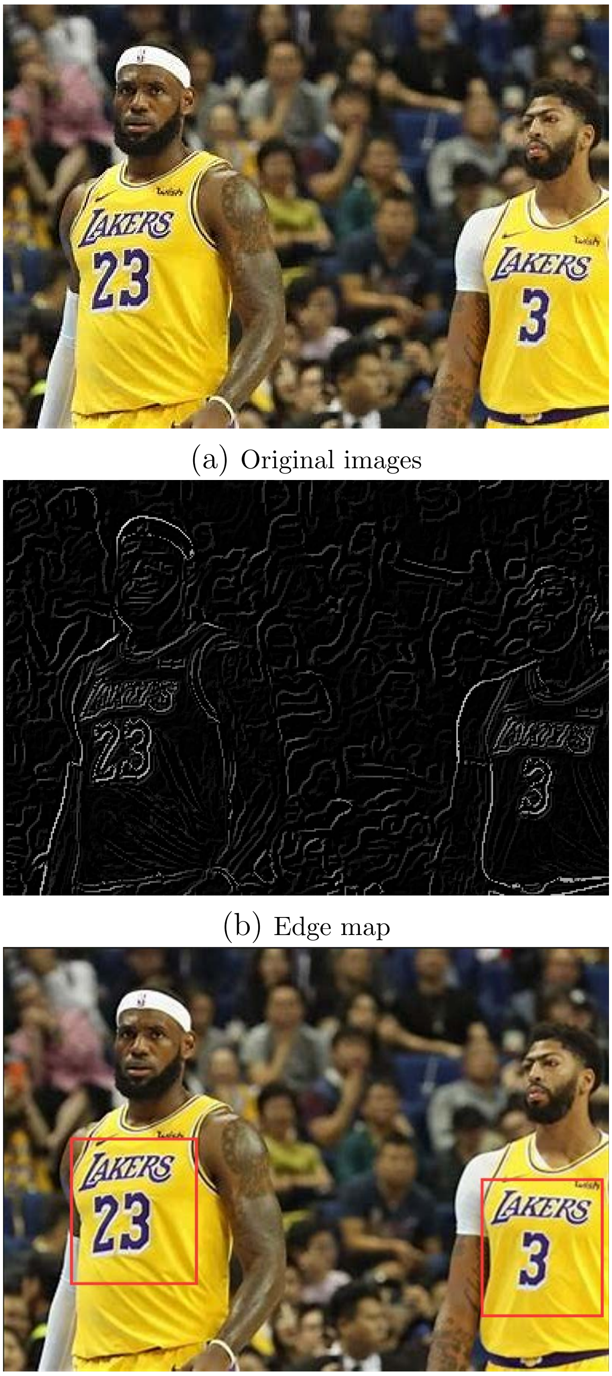
Frame rate	**Number of players**	**Accuracy**	**Recall rate**	**Average time**
400	10	91.2%	90.35%	0.571 s/frame

5.4. **Player number recognition.** Under various complex conditions, there are large differences in the contrast, clarity and brightness of the acquired samples due to the different sample acquisition times and the resolution of the acquisition devices. To validate the proposed improved LeNet-5 is compared with BPNN, LeNet-5 and LSTM and the results are shown in Table 3.

Table 3. Overall test results of different algorithms under complex conditions

Recognition algorithm	**Number of test samples**	**Recognition success**	**Recognition accuracy**
BPNN	200	173	86.5 %
LeNet-5	200	181	90.5 %
LSTM	200	187	93.5 %
Improved LeNet-5	200	191	95.5 %

It can be seen that the improved LeNet-5 has the highest average accuracy of 95.5%. This indicates that this model performs optimally on this task. In contrast, BPNN has



(a) Original images

(b) Edge map

(c) Results of testing

Figure 8. Jersey area positioning results

the lowest average accuracy of 86.5%. This indicates that this model performed the worst on this task. The other two models had average accuracies in between, 90.5% and 93.5%, respectively.

6. **Conclusion.** This work proposes an approach using computer vision and machine learning to detect and recognise players' numbers from sports video sequences. Firstly, the colour component template in HSI space is set so as to extract the colour regions that are similar to the base colour of the jersey, followed by a threshold judgment on the number of candidate regions. If there are too few candidate regions or too many interference regions,

then Gaussian filtering is performed to lay the foundation for later edge feature extraction while maintaining the colour image. Next, edge information detection is performed, and then candidate regions are formed using morphological and custom region connection methods. Finally, HOG feature extraction is performed on the region after the above processing, and the pseudo-jersey region is eliminated using SVM binary classification method to obtain the real jersey region. Shuffle module is introduced in LeNet-5 to improve the performance of the model. The interaction between feature maps can be increased by inserting Shuffle modules between the convolution and pooling layers of LeNet-5. In addition, by adjusting the structure of the module to maximise the retention of important information in the image, the new module is called R-Shuffle. the improved LeNet-5 has the highest average accuracy of 95.5%.

Acknowledgment. This work is supported by Jiangsu Ocean University Undergraduate Education and Teaching reform project (No. JGX2022034), and General Project of Philosophy and Social Science Research in Universities of Jiangsu Province (No. 2023SJYB1824).

REFERENCES

[1] N. Buch, S. A. Velastin, and J. Orwell, "A review of computer vision techniques for the analysis of urban traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 3, pp. 920-939, 2011.

[2] J. Gong, and C. H. Caldas, "Computer vision-based video interpretation model for automated productivity analysis of construction operations," *Journal of Computing in Civil Engineering*, vol. 24, no. 3, pp. 252-263, 2010.

[3] M. Romero, J. Summet, J. Stasko, and G. Abowd, "Viz-A-Vis: Toward visualizing video through computer vision," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1261-1268, 2008.

[4] T. D'Orazio, and M. Leo, "A review of vision-based systems for soccer video analysis," *Pattern Recognition*, vol. 43, no. 8, pp. 2911-2926, 2010.

[5] T. Sayed, M. H. Zaki, and J. Autey, "Automated safety diagnosis of vehicle–bicycle interactions using computer vision analysis," *Safety Science*, vol. 59, pp. 163-172, 2013.

[6] X. Li, et al., "Transfer learning in computer vision tasks: Remember where you come from," *Image and Vision Computing*, vol. 93, 103853, 2020.

[7] G. Thomas, et al., "Computer vision for sports: Current applications and research topics," *Computer Vision and Image Understanding*, vol. 159, pp. 3-18, 2017.

[8] F. Chadebecq, et al., "Computer vision in the surgical operating room," *Visceral Medicine*, vol. 36, no. 6, pp. 456-462, 2020.

[9] B. T. Naik, et al., "A comprehensive review of computer vision in sports: Open issues, future trends and research directions," *Applied Sciences*, vol. 12, no. 9, 4429, 2022.

[10] M. Stein, et al., "Bring it to the pitch: Combining video and movement data to enhance team sport analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 13-22, 2017.

[11] F. Zhang, et al., "Video salient region detection model based on wavelet transform and feature comparison," *EURASIP Journal on Image and Video Processing*, vol. 2019,58, 2019.

[12] F. Zhang, et al., "Human motion recognition based on SVM in VR art media interaction environment," *Human-centric Computing and Information Sciences*, vol. 9, 40, 2019.

[13] R. Zhang, et al., "Multi-camera multi-player tracking with deep player identification in sports video," *Pattern Recognition*, vol. 102, pp. 107260, 2020.

[14] E. Zhou, and H. Zhang, "Human action recognition toward massive-scale sport sceneries based on deep multi-model feature fusion," *Signal Processing: Image Communication*, vol. 84, 115802, 2020.

[15] L. Kong, et al., "A joint framework for athlete tracking and action recognition in sports videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 532-548, 2019.

[16] W.-L. Lu, et al., "Learning to track and identify players from broadcast sports videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1704-1716, 2013.

[17] M. Šaric, et al., "Player number localization and recognition in soccer video using hsv color space and internal contours," *International Journal of Electrical and Computer Engineering*, vol. 2, no. 7, pp. 1408-1412, 2008.

[18] S. Messelodi, and C. M. Modena, "Scene text recognition and tracking to identify athletes in sport videos," *Multimedia Tools and Applications*, vol. 63, no. 2, pp. 521-545, 2013.

[19] M. A. Hameed, et al., "An adaptive image steganography method based on histogram of oriented gradient and PVD-LSB techniques," *IEEE Access*, vol. 7, pp. 185189-185204, 2019.

[20] S. T. Babu, and C. S. Rao, "Efficient detection of copy-move forgery using polar complex exponential transform and gradient direction pattern," *Multimedia Tools and Applications*, vol. 82, no. 7, pp. 10061-10075, 2023.

[21] M. Mohammadi, et al., "A comprehensive survey and taxonomy of the SVM-based intrusion detection systems," *Journal of Network and Computer Applications*, vol. 178, pp. 102983, 2021.

[22] V. K. Chauhan, et al., "Problem formulations and solvers in linear SVM: a review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 803-855, 2019.

[23] J. Zhang, et al., "A novel deep LeNet-5 convolutional neural network model for image recognition," *Computer Science and Information Systems*, vol. 19, no. 3, pp. 1463-1480, 2022.

[24] Y. Fan, et al., "A better way to monitor haze through image based upon the adjusted LeNet-5 CNN model," *Signal, Image and Video Processing*, vol. 14, pp. 455-463, 2020.

[25] Q. Shi, et al., "Shuffle-invariant network for action recognition in videos," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 3, pp. 1-18, 2022.

[26] M. Kamiyama, and A. Taguchi, "Color conversion formula with saturation correction from HSI color space to RGB color space," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 104, no. 7, pp. 1000-1005, 2021.

[27] T. Li, et al., "The image-based analysis and classification of urine sediments using a LeNet-5 neural network," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 8, no. 1, pp. 109-114, 2020.

[28] Z. Cui, et al., "Ship detection in large-scale SAR images via spatial shuffle-group enhance attention," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 379-391, 2020.

[29] F. Zhang, et al., "An effective method for the abnormal monitoring of stage performance based on visual sensor network," *International Journal of Distributed Sensor Networks*, vol. 14, no. 4, pp. 1-11, 2018.

[30] B. Shiroud Heidari, et al., "Optimization of process parameters in plastic injection molding for minimizing the volumetric shrinkage and warpage using radial basis function (RBF) coupled with the k-fold cross validation technique," *Journal of Polymer Engineering*, vol. 39, no. 5, pp. 481-492, 2019.