Recognition of Student Classroom Behaviors Based on Moving Target Detection

Bin Wu, Chunmei Wang*, Wei Huang, Da Huang, Hang Peng

Yueyang Vocational and Technical College, Yueyang 414000, China

Corresponding Author Email: gxrefuly@163.com

ABSTRACT

Classroom teaching, as the basic form of teaching, provides students with an important channel to acquire information and skills. The academic performance of students can be evaluated and predicted objectively based on the data on their classroom behaviors. Considering the complexity of classroom environment, this paper firstly envisages a moving target detection algorithm for student behavior recognition in class. Based on region of interest (ROI) and face tracking, the authors proposed two algorithms to recognize the standing behavior of students in class. Moreover, a recognition algorithm was developed for hand raising in class based on skin color detection. Through experiments, the proposed algorithms were proved as effective in recognition of student classroom behaviors.

1. INTRODUCTION

In recent years, multiple state-of-art techniques have been integrated for behavior recognition, ranging from computer image processing, digital image processing, to behavior pattern analysis. The combined strategies can analyze the target image series deeply, and decipher specific human behaviors from the data features in the images, whether the behaviors involve a single target or multiple targets. So far, moving target tracking and recognition have become relatively mature. However, human behavior recognition has not been widely applied to classroom teaching, owing to the complexity of the classroom environment.

With the continuous progress of education and teaching methods, many education recording and broadcasting systems (ERBSs) have emerged based on intelligent recognition of video images, and penetrated classroom teaching. The intelligence, integrity, and effectiveness of ERBSs hinge on the effective recognition of human behaviors. In the presence of an intelligent ERBS, the activities in class can be judged according to student behaviors in class, facilitating the improvement of the teaching plan.

At present, behavior recognition is mostly realized based on moving target detection, which aims to separate moving targets from the complex background in the original video images. There are roughly two kinds of moving target detection methods: low-level methods, and high-level methods.

The low-level methods are grounded on the acquisition of image information. Moving targets can be detected with low-level image features, such as the speed, contour, and trajectory of foreground target, as well as optical flow. The low-level moving target detection algorithms mainly include Gaussian mixture model (GMM) [1], visual background extraction method (VIBE) [2], and optical flow method [3].

The high-level methods are rooted in body structure analysis. The most representative high-level moving target detection algorithms take the following models as the basis: human body point model [4], two-dimensional (2D) human body model [5], and three-dimensional (3D) human body model [6].

Drawing on the low-level moving target detection methods, this paper attempts to design a recognition method for student behaviors in the complex classroom environment. Based on region of interest (ROI), face tracking, and skin color detection, three algorithms were developed separately to recognize the standing and hand raising behaviors of students in class. The effectiveness of these algorithms was verified through experiments.

2. LITERATURE REVIEW

It is much more complex to recognize behaviors in a series of video images than to decipher behaviors from static images. Quite many scholars have summarized the research into behavior recognition. For example, Cristani et al. [1] reviewed the latest progress and challenges in automatic monitoring of behaviors, from the perspective of social signal processing. Huang et al. [7] summed up the influence of various feature fusion algorithms on behavior recognition accuracy, under the framework of word packet model. Sun et al. [8] combed through the application of artificial feature representation and deep learning in behavior recognition.

In behavior recognition, the useful features can be divided into global features and local features. The global features offer a panoramic picture of human behaviors, that is, describe the human body as a whole. To obtain global features, it is necessary to separate the human body from the video through target tracking or foreground-background separation. Through Kalman filtering, Al-Shabi et al. [9] realized the smooth estimation of actor's 3D pose, and described human behaviors in image series. Cheng et al. [10] extracted a contour sequence from the series of video images, and combined frame differences to illustrate the behaviors in historical video images on target motion. Baloch and Krim [11] assigned a value to each point inside the contour to indicate the mean time of random walk from the point to the edge, and reliably

extracted various shape features, including partial structure and rough skeleton, by solving Poisson equation.

Unlike global features, local features focus on local areas that catch visual attention, such as corners and edges. Wu et al. [12] extended Harris detector from corner detection in static images to the 3D space time. Drawing on the scale space theory, Yang et al. [13] took the determinant of Hessian matrix as the saliency measure, proposed a spatiotemporal representation method of spatiotemporal-invariant interest points, and covered the whole video with dense interest points. Xu et al. [14] introduced dense sampling method for the first time to static image interpretation: the image features were extracted after dividing the image into several blocks of different scales. Wang et al. [15] pointed out the sharp differences between 2D space domain and time domain of videos, and considered tracking the feature points in time domain a better choice than detecting spatiotemporal interest points. Based on low-level information of optical flow, Ou and Sun [16] proposed a representation method for middle-level motion features, which focus on the local areas of image series, and adopted AdaBoost method to enhance the discriminability of the features between behavior classes, making them easier to calculate in real time. Basly et al. [17] represented the behaviors in a series of video images with a middle-level component, which has a continuous spatial structure and coherent motion features. Considering the obvious semantics of high-level features, Neviarouskaya et al. [18] put forward a behavior recognition method based on attitude, which outperforms underlying feature-based method in the presence of massive noises. Fyffe et al. [19] modeled the behaviors in video as a set of partial key poses, that is, the time series of local key frames, and thus located the behaviors in real video streams.

3. CLASSROOM VIDEO IMAGE PREPROCESSING BASED ON THRESHOLDING AND BOUNDARY TRACKING

Owing to the complexity of the classroom environment, the preprocessing of video images directly bears on whether the student behaviors can be identified effectively. Since the original data are in the form of video, it is necessary to segment the video file into multiple frames, and process each frame, preparing for the subsequent behavior recognition. After all, the essence of video processing is image processing.

The frame rate is an important parameter in image processing. It refers to the number of images that can be transmitted in a second, that is, the number of times that the image processor can refresh per second. Normally, the frame rate is measured in the unit of frames per second (fps). This paper sets the frame rate to 25 fps to enhance the clarity of behavior recognition and eliminate redundant data. Figure 1 shows the flow of video sampling in preprocessing.
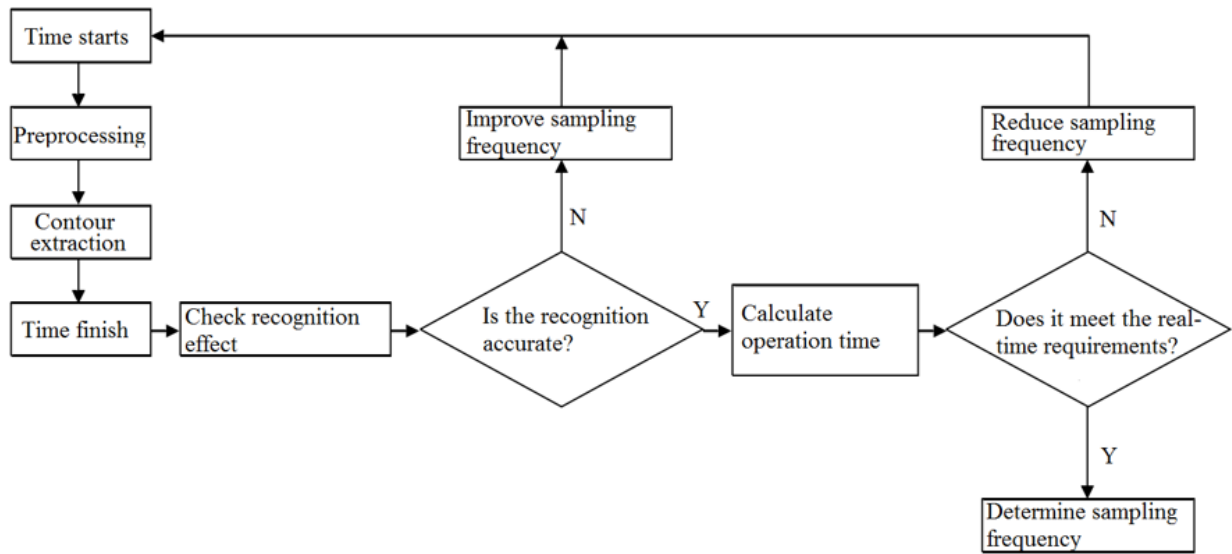

Figure 1. Flow of video sampling in preprocessing

The original video is in true color. Thus, the images segmented from the video are color images, which need to be grayed and binarized. In addition, denoising should be performed on the images to remove the interreference of instruments and transmission distortion.

Color space conversion is a common operation in digital image processing. The most common coordinate system for color space is the red-green-blue (RGB) space coordinate system. However, the RGB space can only reflect the true color, failing to represent the other parameters of the image. To solve the problem, the RGB color space is often converted into hue-saturation-intensity (HSI) color space by:

$$H = \begin{cases} \theta, & B \leq G \\ 360 - \theta, & B > G \end{cases} \tag{1}$$

$$S = 1 - 3/(R + G + B) \times [min (R, G, B)] \tag{2}$$

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

where, $\theta = \arccos \frac{\frac{1}{2}[(R-G)+(R-B)]}{[(R-G)^2+(R-B)(G-B)]^{\frac{1}{2}}}$.

Color mainly plays an ornamental role, without containing any information of the image. Direct processing of color images brings a huge computing load, dragging down the processing speed. Thus, graying was introduced to reduce the spatial complexity and information volume of the images, making it possible to improve processing speed and recognition efficiency.

This paper chooses the weighted average method [20] for graying. Specifically, different weights were assigned to R, G, and B components, and the weighted average was computed as the pixel value in the gray image:

$$Y = 0.299 \times R + 0.578 \times G + 0.114 \times B \tag{4}$$

The weights obtained by the weighted average method are in line with human visual experience, because they are based on the sensitivity of human eyes to color.

Next, mean filtering [21], a typical linear filtering algorithm was applied on the gray images. Let (s, t) be a pixel randomly selected from an image I (x, y), R be the neighborhood around the pixel, and P be the number of pixels in the neighborhood. Then, the gray value of pixel (s, t) can be obtained by summing and averaging the gray values of the pixel and its neighborhood pixels. With pixel (s, t) at the center, a 3×3 filter template was defined. Then, the pixel value of (s, t) can be calculated by:

$$I(s,t) = 1/9 \sum_{i=-1}^{1} \sum_{j=-1}^{1} I(s+i, t+j) \tag{5}$$

Suppose noise n (x, y) is uncorrelated in the space, with an expectation of 0 and a variance of σ^2. Let o (x, y) be the noise-free image. Then, the noisy image can be processed by mean filtering:

$$I(s,t) = 1/P \sum I(i,j) = 1/P \sum o(i,j) + 1/P \sum n(i,j) \tag{6}$$

After mean filtering, the average value of the noise remains unchanged, but the variance turns into $\frac{1}{P}\sigma^2$, suggesting that the noise has been suppressed.

As mentioned before, this paper mainly deals with student behaviors in complex classroom environment. To acquire the features of student behaviors, it is important to separate the moving targets from the background through image thresholding. The key of image thresholding is to find a suitable threshold to differentiate the gray value of the moving target from that of the background. Such a threshold can convert the gray image into a binary image, with a sharp contrast between the moving target and the background.

Maximum inter-class variance method [22] is the most popular way to determine the threshold. Based on the maximum variance between foreground and background, this method can find the best threshold to divide an image into foreground and background.

Let $w \times h$ be the size of image; $0 \sim 255$ be the range of gray value; n_i be the number of gray values i in the image. Then, the total number of pixels in the image equals the sum of the number of pixels occupied by each gray value:

$$N = \sum_{i=0}^{255} n_i \tag{7}$$

The proportion of each gray value can be defined as:

$$p_i = n_i / N \tag{8}$$

where, $p_i \geq 0$ $(\sum_{i=0}^{255} p_i = 1)$.

Let K be the threshold to divide the image into part A and part B. The former part contains the pixels with gray value in [0, K], and the latter contains the pixels with gray value in [K+1, K-1]. Then, the proportion of A and B can be

respectively calculated as:

$$\rho_1 = \sum_{i=0}^{K} n_i / N \tag{9}$$

$$\rho_2 = \sum_{K}^{255} n_i / N = 1 - \rho_1 \tag{10}$$

The mean gray value of A and B can be obtained as:

$$u_1 = \sum_{i=0}^{K} i \times p_i / \rho_1 \tag{11}$$

$$u_2 = \sum_{i=K+1}^{255} i \times p_i / \rho_2 \tag{12}$$

The mean gray value of the whole image can be derived by:

$$u = \rho_1 u_1 + \rho_2 u_2 \tag{13}$$

With K as the threshold for image segmentation, the variance between A and B can be described as:

$$\sigma^2 = \rho_1(u_1 - u)^2 + \rho_2(u_2 - u)^2 = \rho_1 \rho_2 (u_1 - u_2)^2 \tag{14}$$

The variance was calculated continuously as the threshold K changed from 0 to 255. The K value corresponding to the maximum variance was taken as the threshold for image segmentation.

Hence, video image preprocessing is completed through color space conversion, image graying, image denoising, and image thresholding, laying the basis for the design of behavior recognition algorithms.

4. DESIGN OF STANDING BEHAVIOR RECOGNITION ALGORITHMS

The standing behavior of students in class has two prominent features: the upper body must move significantly; the subject of the behavior must be a person, not any other object. According to the two features, this paper puts forward two recognition methods for this behavior. The former method is based on ROI recognition, and the latter on face tracking.

4.1 ROI-based standing behavior recognition method

In classroom environment, a student standing up must be taller than he/she sitting down. Thus, height was adopted as an indicator of the standing behavior in class. When a student stands up, the largest change in pixels occurs in the head, while the pixels below the head change little and disorderly. During the recognition on standing behaviors, the variation in the lower part of the student will seriously interfere with the analysis. Therefore, the ROI was cropped from the original video images in the following steps:

Step 1. Set the cropping height of the image by method of bisection: $h = (h_{max} + h_{min})/2$.

Step 2. Perform inter-frame subtraction between the first two frames in the video, according to the pixel values in the upper and lower parts defined by the height in Step 1.

Step 3. If the pixel difference between the upper and lower parts is nonzero, the cropping height is too low (the lower part also moves); in this case, adjust the cropping height by method of bisection: $h' = (h + h_{max})/2$. If the pixel difference is zero, the cropping height is too high (the upper part is not the student); in this case, adjust the cropping height by method of

bisection: $h' = (h + h_{min})/2$. Repeat this process until the height difference $\Delta h \leq 1$. Then, take the region between (h', h_{max}) as the ROI.

After that, no part in the image other than ROI needs to be analyzed. Here, background subtraction is performed to remove the background from every cropped frames. Specifically, the pixel value of the current frame was compared against that of the background image to detect motions. The crux of this method is to build a robust background model, which can adapt to illumination changes, small target motions in the background, and noise influence. The background subtraction method can be defined as:

$$B_t(x,y) = \begin{cases} 255, & |I_t(x,y) - G_t(x,y) \geq T_t| \\ 0, & |I_t(x,y) - G_t(x,y) < T_t| \end{cases} \quad (15)$$

where, $I_t(x,y)$ is the pixel value of the current frame; $G_t(x,y)$ is the pixel value of the background images; $B_t(x,y)$ is the difference image between the current frame and background image; t is the number of frames; T is the threshold.

Figure 2 provides an example of standing behavior recognition after background subtraction. It can be seen that, thanks to background subtraction, the head motion area of the standing student was effectively detected.
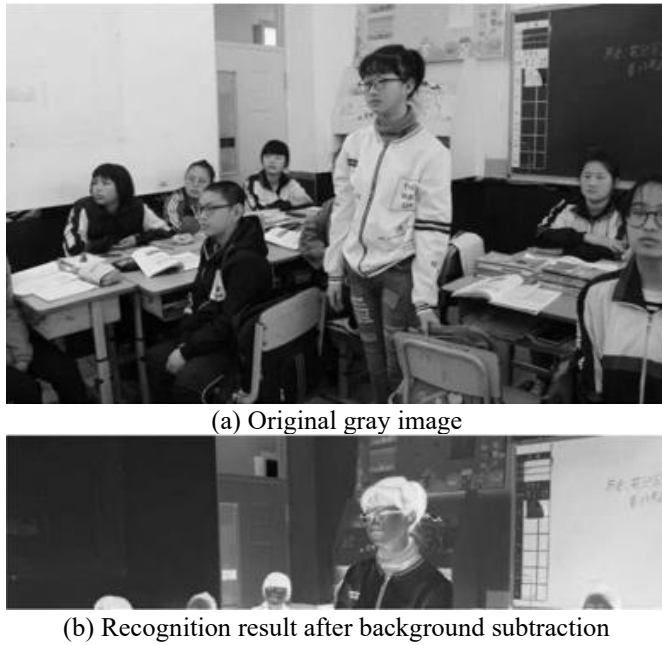
(a) Original gray image


(b) Recognition result after background subtraction

Figure 2. An example of standing behavior recognition after background subtraction

4.2 Face tracking-based standing behavior recognition method

Face tracking is a way to improve the accuracy of face detection. Because all students face the same directly in class, all the faces in the video images must be directly forward. Thus, the Viola Jones (VJ) algorithm [23] was employed to detect faces in video images. This algorithm involves four stages: Haar feature extraction, integral image creation, AdaBoost training, and cascade classification.

Haar feature extraction aims to find simple rectangular features based on pixel intensity. These features reflect the light-dark relationship between local areas of the image, and provide effective information for object differentiation. For face detection, only four basic structures, i.e., Haar features, need to be obtained through feature extraction (Figure 3). Through scaling, rotation, or translation, these basic structure can be transformed into a series of similar structures.

(a) Edge feature (b) Linear feature (c) Diagonal feature (d) Central feature
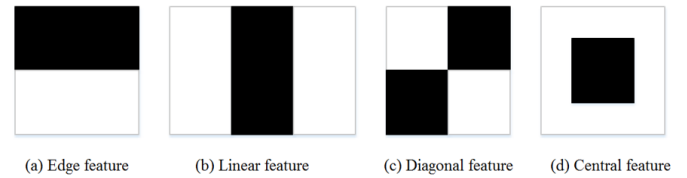
Figure 3. Haar features

To calculate Haar features, it is necessary to find the sum of all pixels in the box. However, the computing load would be too large if every pixel is traversed to obtain the sum. The VJ algorithm offers a new data structure for the summation: creating an integral image of all pixels:

$$I(x,y) = \sum_{x' \leq x} \sum_{y' \leq y} f(x', y') \quad (16)$$

where, x, y, x', and y' are the coordinates of every pixel.

Then, AdaBoost, which can perform feature selection and classifier training at the same time, was called to synthetize a subset of relevant features into an effective classifier for face detection. To save time, the detector should focus on the potential areas of positive samples, rather than traverse all the areas by all the features. Hence, a cascade classifier was designed to remove most negative areas with a small number of features, and then remove the other non-positive areas with complex features.

Figure 4 presents the result of face detection by VJ algorithm. Obviously, every face looking into the camera was detected by the algorithm.
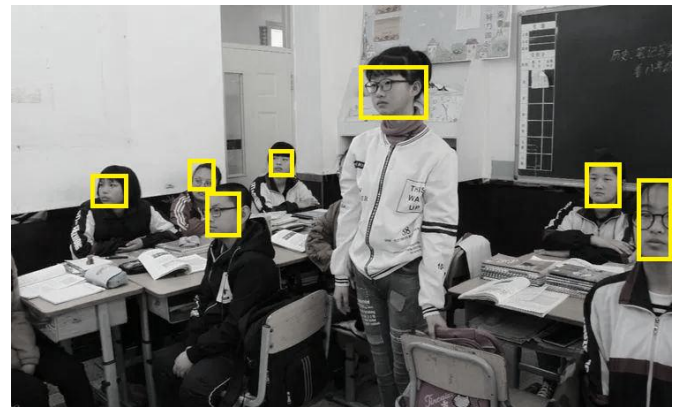
Figure 4. Face detection result of VJ algorithm

Location of faces is critical to the accurate tracking and intuitive, simple expression of standing behaviors. Thus, the face blocks in the image were acquired during feature extraction. The inter-frame coherence was highlighted in behavior analysis, due to the continuity of target motions.

First, the two most similar adjacent frames were identified with a particular index. In the class environment, the students sit in different places. Normally, the students can be located by the center of the image. However, when two students sit in the front and back rows, it would be hard to differentiate between the coordinates of their positions. To overcome the difficulty, this paper introduces the block area to help

determine the coherence between consecutive blocks, and then track the standing students. Figure 5 shows the face tracking results on Figure 4. It can be seen that Student 5 stood up in the class.
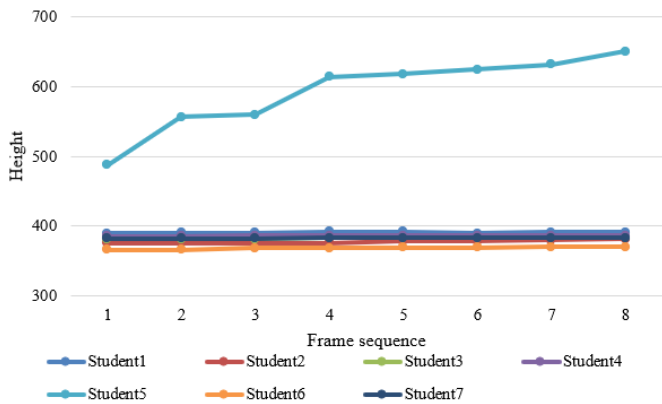

Figure 5. Face tracking results

5. DESIGN OF SKIN COLOR-BASED HAND RAISING RECOGNITION ALGORITHM

During the class, the students might raise their hands in different forms. It is difficult to tell if they are raising hands by the shape or height of their hands. Thus, this paper chooses to determine the hand raising behavior with the only constant factor: the skin color of their hands.

Through experiments, the skin color range was determined for the specific color space, and an operable skin color model was established, making skin color the core of hand raising recognition. Table 1 shows the hand areas with different yet typical skin colors and background illuminations, which were obtained through repeated tests. Table 2 shows the HSI values corresponding to the RGB values.

Table 1. RGB values of skin areas

Skin area	R	G	B
1	252.6	252.3	252.2
2	243.5	240.8	233.7
3	238.2	222.8	216.9
4	230.2	209.6	218.7
5	189.4	163.7	202.7
6	156.3	128.5	139.7

Table 2. HSI values of skin areas

Skin area	H	S	I
1	6.23	0.102	248.9
2	6.19	0.151	220.7
3	6.02	0.226	201.1
4	5.96	0.258	168.9
5	5.84	0.367	155.6
6	5.77	0.392	138.5

From the above data, it can be seen that the HSI values of all skin areas have certain thresholds:

$$5.77 \leq H \leq 6.23$$
$$0.10 \leq S \leq 0.39 \qquad (17)$$
$$139 \leq I \leq 249$$

Then, the skin areas can be identified rather accurately by superposing the two criteria (RGB + HSI). But faces are often included in the detection results on skin colors. It is important to remove the face areas, the largest interference in hand raising recognition. This paper adopts the scan line seed fill algorithm to recognize the face areas, taking advantage of the fact that raising hands are usually long strips, while the face is round with dark holes. The face areas were recognized in the following steps:

Step 1. Select a point (x, y) inside the boundary of the region as the seed point.

Step 2. Starting from the current seed point (x, y), scan along the left and right directions ((x-1, y) and (x+1), y) of the scan line y, and then scan along the upper and lower directions in turn until encountering the point with no pixel value in the boundary. Then, take the point as a new seed point (x_n, y_n). If there is no point with no pixel value in the boundary, then there is no hole in the connected domain, and the algorithm terminates.

Step 3. Start cleaning from the current seed point, first along the upper and lower directions of the scan line y_n, and then along the left and right directions of the scan line y, until reaching each boundary point of the region.

Figure 6 shows the hand raising behaviors identified by the proposed algorithm. It can be seen that the recognition was not affected by faces.

Figure 6. Results of hand raising recognition algorithm

6. CONCLUSIONS

This paper applies several state-of-the-art image processing methods to classroom behavior recognition, and recognizes the behaviors of standing up and hand raising in the classroom environment. Specifically, the collected video was preprocessed through color space conversion, graying, denoising, and thresholding. Then, several algorithms were designed to recognize the standing up and hand raising behaviors of students in class. Experimental results show that our algorithms work effectively on behavior recognition in complex classroom environment.

REFERENCES

[1] Cristani, M., Raghavendra, R., Del Bue, A., Murino, V. (2013). Human behavior analysis in video surveillance: A social signal processing perspective. Neurocomputing,

100: 86-97. https://doi.org/10.1016/j.neucom.2011.12.038

[2] Guang, H., Wang, J., Xi, C. (2014). Improved visual background extractor using an adaptive distance threshold. Journal of Electronic Imaging, 23(6): 063005. https://doi.org/10.1117/1.JEI.23.6.063005

[3] Han, P., Du, J., Zhou, J., Zhu, S. (2013). An object detection method using wavelet optical flow and hybrid linear-nonlinear classifier. Mathematical Problems in Engineering. https://doi.org/10.1155/2013/965419

[4] Zhao, X., Satoh, Y., Takauji, H., Kaneko, S.I., Iwata, K., Ozaki, R. (2011). Object detection based on a robust and accurate statistical multi-point-pair model. Pattern Recognition, 44(6): 1296-1311. https://doi.org/10.1016/j.patcog.2010.11.022

[5] Kang, B., Zhu, W.P. (2015). Robust moving object detection using compressed sensing. IET Image Processing, 9(9): 811-819. https://doi.org/10.1049/iet-ipr.2015.0103

[6] Wang, X., Ning, C., Xu, L. (2015). Spatiotemporal saliency model for small moving object detection in infrared videos. Infrared Physics & Technology, 69: 111-117. https://doi.org/10.1016/j.infrared.2015.01.018

[7] Huang, S., Jin, L., Wei, X. (2011). Online heterogeneous feature fusion for visual recognition. 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, pp. 799-803. https://doi.org/10.1109/ICDMW.2011.131.

[8] Sun, G.X., Bin, S., Dong L.B. (2018). New trend of image recognition and feature extraction technology introduction. Traitement du Signal, 35(3-4): 205-208.

[9] Al-Shabi, M., Gadsden, S.A., Habibi, S.R. (2013). Kalman filtering strategies utilizing the chattering effects of the smooth variable structure filter. Signal Processing, 93(2): 420-431. https://doi.org/10.1016/j.sigpro.2012.07.036

[10] Cheng, F.C., Huang, S.C., Ruan, S.J. (2011). Illumination-sensitive background modeling approach for accurate moving object detection. IEEE Transactions on Broadcasting, 57(4): 794-801. https://doi.org/10.1109/TBC.2011.2160106

[11] Baloch, S.H., Krim, H. (2007). Flexible skew-symmetric shape model for shape representation, classification, and sampling. IEEE Transactions on Image Processing, 16(2): 317-328. https://doi.org/10.1109/TIP.2006.888348

[12] Wu, Z., Li, J., Guan, Z., Yang, H. (2016). Obstacle regions extraction method for unmanned aerial vehicles based on space–time tensor descriptor. Journal of Electronic Imaging, 25(5): 053029. https://doi.org/10.1117/1.JEI.25.5.053029

[13] Yang, B.J., Bahramy, M.S., Arita, R., Isobe, H., Moon, E.G., Nagaosa, N. (2013). Theory of topological quantum phase transitions in 3D noncentrosymmetric systems. Physical Review Letters, 110(8): 086402. https://doi.org/10.1103/PhysRevLett.110.086402

[14] Xu, Z., Hu, R., Chen, J., Chen, C., Chen, H., Li, H., Sun, Q. (2017). Action recognition by saliency-based dense sampling. Neurocomputing, 236: 82-92. https://doi.org/10.1016/j.neucom.2016.09.106

[15] Wang, J., Fan, R. (2003). Two-dimensional time-domain volume integral equations for scattering of inhomogeneous objects. Radio Science, 38(4): 9-1-9-12. https://doi.org/10.1029/2000RS002605

[16] Ou, H., Sun, J. (2021). The multidimensional motion features of spatial depth feature maps: An effective motion information representation method for video-based action recognition. Mathematical Problems in Engineering. https://doi.org/10.1155/2021/6670087

[17] Basly, H., Ouarda, W., Sayadi, F.E., Ouni, B., Alimi, A.M. (2021). DTR-HAR: Deep temporal residual representation for human activity recognition. The Visual Computer. https://doi.org/10.1007/s00371-021-02064-y

[18] Neviarouskaya, A., Prendinger, H., Ishizuka, M. (2015). Attitude sensing in text based on a compositional linguistic approach. Computational Intelligence, 31(2): 256-300. https://doi.org/10.1111/coin.12020

[19] Fyffe, G., Jones, A., Alexander, O., Ichikari, R., Debevec, P. (2014). Driving high-resolution facial scans with video performance capture. ACM Transactions on Graphics (TOG), 34(1): 1-14. https://doi.org/10.1145/2638549

[20] Hasanzadeh, N., Forghani, Y. (2019). Improving the accuracy of M-distance based nearest neighbor recommendation system by using ratings variance. Ingénierie des Systèmes d'Information, 24(2): 131-137. https://doi.org/10.18280/isi.240201

[21] Rodríguez, R., Suarez, A.G., Sossa, J.H. (2011). A segmentation algorithm based on an iterative computation of the mean shift filtering. Journal of Intelligent & Robotic Systems, 63(3): 447-463. https://doi.org/10.1007/s10846-010-9503-y

[22] Wang, Y.N., Yang, Y.M., Zhang, P.Y. (2020). Gesture feature extraction and recognition based on image processing. Traitement du Signal, 37(5): 873-880. https://doi.org/10.18280/ts.370521

[23] Chirra, V.R.R., Uyyala, S.R., Kolli, V.K.K. (2019). Deep CNN: A machine learning approach for driver drowsiness detection based on eye state. Revue d'Intelligence Artificielle, 33(6): 461-466. https://doi.org/10.18280/ria.330609