OPEN

Confusion matrix and minimum cross-entropy metrics based motion recognition system in the classroom

Ming-Te Wu

This research proposes a motion recognition system for early detection of students' physical aggressive behavior in the classroom. The motion recognition system recognizes physical attacks so that teachers can resolve disputes early to reduce other greater injuries. In the beginning, cameras were used in this system to monitor students' classroom activities and to obtain body images by removing background and saliency maps. Two angles from arm to shoulder and shoulder to the center of the body are then measured and the velocity between the two frames from the movement of the body is detected, and use these angle and velocity values as the criterion for judging whether it is a physical attack. In the end, the accuracy of the proposed algorithms is verified by using the confusion matrix based on machine learning and the minimum cross entropy based on neural networks. The simulation proves that the proposed algorithm can correctly detect the attack behavior of the collected videos.

The report by Casas¹ pointed out that regardless of whether the child uses indirect or direct physical aggression, there is a significant correlation between the parent's parenting style, attachment relationship, and psychological control behavior. The findings emphasize the importance of indirect and physical aggression and parenting behaviors of young children, as well as the potential connection between the child's gender and the mother and father who raises the child. Kupersmidt² showed that in the field of early childhood education, teachers estimated that 10% of preschool children have daily bullying behaviors. Aggressive behaviors appear in early childhood³, and even these early forms of aggressive behavior will always exist and eventually become a social problem⁴. Scholars' research found that⁵ teachers assessing children's aggressive behavior at the age of 8 are related to maladjustment in school in the future and lead to long-term unemployment in adulthood. Whenever aggressive behavior occurs, people almost always have habitual gestures. If a system can be developed that can detect aggressive gestures as soon as possible, it can reduce the occurrence of injuries. Many articles on gesture research have been published recently^{6–9}. Kim⁶ developed a system that can change the 3-D spatial coordinates of each joint to the 2-D angle representation of the corresponding arm, and display the specific time pattern of each dynamic gesture by using a discrete hidden Markov model. A hand-raising detection system initially obtains motion data through time difference and then uses a threshold to obtain the object of interest proposed by Yao⁷. To realize long-distance human-computer interaction, Kim⁸ and Lupinetti¹⁰ developed an arm gesture detection system. The system initially recognizes the user's face and arms through the background removal method and then determines which gesture is based on the position where the arm is raised. An arm pose segmentation method without monitoring was proposed by Simão⁹. This method is based on the use of thresholds, which can divide the input continuous flow of information, including unsegmented and unbounded, into dynamic and static segmented types. Although these methods can detect gestures, most of them use thresholds as a solution to detect gestures. The threshold used to detect gestures can usually get good results, but if it is applied to attacks between people, the accuracy is insufficient. Akl^{11,12} published a gesture detection technology that uses data obtained from a 3-axis accelerometer, including the training phase and the test phase¹³. A study uses the three gesture detection technologies of artificial neural network and dynamic time warp, hidden Markov model at the same time, and compares these three methods with the accelerometer in the mobile equipment, and then finds that the best approach is the dynamic time warping¹³. To model the image characteristics of the surface electromyography signal, Tsinganos¹⁴ applied the Hilbert space-filling curve and classified it through a

Department of Information Technology, Kao Yuan University, No. 1821, Zhongshan Rd., Luzhu Dist., Kaohsiung 82151, Taiwan. email: t90078@cc.kyu.edu.tw

convolutional neural network. In recent years, other related articles have been put forward one after another^{15–18}. Besides, there are many relevant and interesting studies for reference^{19–21}. To improve the accuracy of detecting gestures, the above methods use artificial neural networks and convolutional neural networks. These methods have indeed improved the recognition ability, but there is still room for improvement in accuracy. Occasionally, conflicts between students on campus cause mutual injuries. This kind of scene happens every year. Therefore, this motivates us to develop a motion recognition system that can detect physical conflicts as early as possible and immediately eliminate them to reduce serious injuries. Since the matrix is a very effective and popular modeling tool in various applications, some articles^{22–26} have been proposed and applied in other fields in recent years.

A scheme for identifying aggressive behaviors is proposed in this research and teachers can be notified immediately. When there are dangerous aggressive behaviors among students, teachers can eliminate conflicts as soon as possible to prevent aggressive behaviors. This technique uses some cameras to observe students' behavior in class, then background elimination algorithms and saliency map technologies are used to capture the body's position. Next, the angle between the line formed from the arm to the shoulder and the line formed from the shoulder to the center of the body, and the velocity of movement of the center of the body are calculated. Whether the recognized body movement is an aggressive behavior is determined by the angle and velocity values in this algorithm. In the end, confusion matrix and minimum cross-entropy are applied, based on machine learning and neural network, to verify the accuracy of the proposed algorithm.

The rest of this article is divided into the following. The background of well-known gesture detection and the proposed method is described in “Background” and “The proposed motion recognition system”. “Experimental results” shows the simulation results. The conclusion of this manuscript is in “Conclusions”.

Background

In this section, a visual attention system, confusion matrix, and cross entropy minimization are introduced as follows. “Visual attention system” presents the visual attention system, “Confusion matrix” introduces the confusion matrix, and “Cross entropy minimization” statements the cross entropy minimization used in this work.

Visual attention system. To capture the region of interest (ROI), time segmentation is often used in video processing⁷. The difference image is calculated by subtracting each pixel in the current frame and the previous frame, namely,

$$D(i, j) = F_t(i, j) - F_{t-1}(i, j) \quad (1)$$

where $D(i, j)$ represents the difference between the two images, and $F_{t-1}(i, j)$ and $F_t(i, j)$ represent the previous frame and the current frame, respectively. The background information is deleted after subtracting two consecutive frames. Image changes in nature are easily affected by time segmentation methods, and a common influencing factor is the non-static background environment. It is a very useful method to eliminate the non-static background environment by using a threshold to check. Yeh¹⁵ define the pixel values of the same offset in two consecutive frames as static pixels and expressed as

$$\text{if } ||D(i, j)|| < \text{Threshold, then } D(i, j) \in S(i, j) \quad (2)$$

where $S(i, j)$ represents a group of static pixels. To find the region of interest, a visual attention system was proposed by Chen¹⁷. In the beginning, color quantization is used in this visual attention system to smooth the color in the texture area. The saliency map is then generated by using the transform color space method. A content-based saliency map is finally formed to calculate the contrast value. In visual attention analysis, the advantage of a content-based saliency map is that it can provide texture, edge intensity, color, and contrast information. In addition, the saliency map scheme is used to extract the region of interest. To highlight the color of the region of interest and smooth texture region, initially, a frame is divided into 4×4 blocks using color quantization in this method to avoid fragile areas due to texture. The value in the RGB color space field is then converted to the value in the XYZ color space field. By using the color space conversion scheme¹⁸, the expression is

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.412 & 0.357 & 0.18 \\ 0.212 & 0.715 & 0.072 \\ 0.019 & 0.119 & 0.95 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \quad (3)$$

Next, the conversion from XYZ color space to LUV color space continues in the saliency map method, and the transformation rows.

$$L = \begin{cases} 116 \sqrt[3]{y_r} - 16y_r > \varepsilon \\ ky_r y_r \leq \varepsilon \end{cases} \quad (4)$$

$$u = 13L(u' - u'_r) \quad (5)$$

$$v = 13L(v' - v'_r) \quad (6)$$

where $y_r = \frac{Y}{Y_r}$, $\varepsilon = 0.008856$, $k = 903.3$,

$$u' = \frac{4X}{X + 15Y + 3Z} \quad (7)$$

$$v' = \frac{9Y}{X + 15Y + 3Z} \quad (8)$$

$$u'_r = \frac{4X_r}{X_r + 15Y_r + 3Z_r} \quad (9)$$

$$v'_r = \frac{9Y_r}{X_r + 15Y_r + 3Z_r} \quad (10)$$

In the end, the contrast values are computed by the obtained frame value from the color quantization and color space conversion. Assuming that there is one pixel for each perceptual unit, then an $N \times M$ size frame has $N \times M$ perceptual units. By the following equation, in other words, the contrast value $C_{i,j}$ is calculated.

$$C_{i,j} = \sum_{q \in A} e(p_{i,j}, q) \quad (11)$$

where q represents the perception unit and $p_{i,j} (i \in [0, N], j \in [0, M])$. The term e is expressed as the Euclidean distance between $p_{i,j}$ and q , and A represents the area surrounding the perceived position (i, j) . The area of A can be determined by adjusting the sensitivity of the perception field. By reducing the size of A , the perceptual field can be made more sensitive. For normalization, the contrast value $C_{i,j}$ is set between 0 and 255 in this article. As long as the contrast value in the saliency map is used as the density, the center point in the saliency map can be regarded as the center of visual attention. The center of visual attention is therefore calculated

$$\begin{cases} x_c = \frac{1}{C_T} \sum_{j=0}^{N-1} C_{i,j} \times i \\ y_c = \frac{1}{C_T} \sum_{i=0}^{M-1} C_{i,j} \times j \end{cases} \quad (12)$$

where $C_T = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} C_{i,j}$.

Confusion matrix. Fawcett¹⁹, Powers²⁰, and Stehman²¹ used the confusion matrix method to improve the accuracy of judging whether it is an attack. The confusion matrix can be illustrated in the table below.

Video	Total frames	True condition			
		Positive	Negative		
Predicted outcome	Positive	True positive (TP)	False positive (FP)	Positive predictive value (PPV) $\frac{TP}{TP+FP}$	False discovery rate (FDR) $\frac{FP}{TP+FP}$
	Negative	False negative (FN)	True negative (TN)	False omission Rate (FOR) $\frac{FN}{FN+TN}$	Negative predictive value (NPV) $\frac{TN}{FN+TN}$
		True positive rate (TPR) Sensitivity $\frac{TP}{TP+FN}$	False positive rate (FPR) $\frac{FP}{FP+TN}$		
	Accuracy $\frac{TP+TN}{T}$	False negative rate (FNR) $\frac{FN}{TP+FN}$	True negative rate (TNR) $\frac{TN}{FP+TN}$		

TP (True Positive) indicates that "predicted outcome is positive" is the same as "true condition is positive." FP (False Positive) means that "the predicted outcome is negative" is different from "true condition is positive." TN (True Negative) means that "predicted outcome is negative" is the same as "true condition is negative." FN (False negative) means that "predicted outcome is negative" is different from "true condition is positive." T (Total Population) represents the number of predictions for all frames. Therefore, the accuracy can be calculated

$$A = \frac{TP + TN}{T} \quad (13)$$

Cross entropy minimization. There is an important function in the neural network learning process that can affect the quality of the model, called the loss function. The focus of the loss function is to calculate the gap between the output value and the actual value; when the gap between the output value and the actual value is larger, the value of the loss function is larger, and vice versa; therefore, the neural network has an important point in the learning process. The purpose of is to minimize the loss function to achieve better classification or prediction results.

The cross entropy function is mainly to evaluate whether the output value and the actual value are very different. Cross entropy is a loss function for probability. Because it can effectively quantify the difference between predicted probability and actual probability, it is often used in classification problems. The formula is as follows,

Figure 1. A general overview of the proposed motion recognition system.

$$-\sum_i Y_i \ln(y_i) \quad (14)$$

Among them, y_i is the value of probability, and Y_i is the probability of the actual category.

The proposed system with motion recognition ability was introduced below. The proposed system initially removes the background of the image and applies a saliency map scheme to extract the ROI part of the body. The proposed algorithms then use the movement speed of the body and the angle between the detected arm and the body in two consecutive frames to determine whether the detected motion is aggressive. Finally, this research exploits the confusion matrix and the minimized cross-entropy from the neural network-like basis, which is detailed in the next section.

The proposed motion recognition system

Figure 1 shows a block diagram of the proposed system with motion recognition. Two stages are provided in the overview: the first stage is the block diagram above the dotted line, which is also the testing phase; the second stage is the block diagram below the dotted line, which is the training phase. In addition, five parts are included in the testing stage. Framing the captured video is the main work of the first part. Second, the comparison of two consecutive frames is performed in the second part. The removal of background and uninteresting background is in the third part. To get the desired ROI, the saliency map motion estimation is implemented in the fourth part. The last part detects the angle of the arm by the angle between the arm to the shoulder and the shoulder to the center of the body, as well as the velocity of the ROI. The training phase contains only one part, that is, by minimizing the cross-entropy and confusion matrix to compute the accuracy of the offensive behavior obtained by the candidate in the testing phase. To make this motion recognition system have the purpose of self-upgrading, the method of improving the accuracy in this paper is to minimize the cross-entropy and adjust the threshold of the ROI movement velocity. The detailed statement is as follows.

The position of the four cameras in the left, right, front, and rear of the classroom is intended to prevent the possibility of gestures being blocked by objects or people. Initially, the background removal is performed on the camera frame using Eqs. (1) and (2), and then the frame is color quantized, and then Eqs. (3) and (4) are used to calculate the color space transformation and contrast values, respectively. Therefore, a saliency map is acquired.

Through Eq. (12), the center point of the ROI object is calculated, and the center position of the ROI in each frame is recorded at the same time after the saliency map is obtained. The center point coordinates of the ROI object, taking the first frame of Fig. 2 as an example, can be calculated as (5,4) and (16,4) by Eq. (12). The ROI object of two consecutive frames is used and the motion velocity of the object is calculated to determine whether the candidate's aggressive behavior is about to occur. Judging that an attack is about to happen is if the movement speed exceeds the threshold. Assuming that the center coordinates of the previous frame F_{t-1} and the current frame F_t are (x_{t-1}, y_{t-1}) and (x_t, y_t) respectively, then the velocity $V_{t-1} = (x_t - x_{t-1}, y_t - y_{t-1})$. Whether the candidate's behavior is offensive is determined by whether the velocity V_{t-1} of the current frame and the velocity V_t of the next frame is greater than the threshold. The two successive velocities of the object are $V_{t-1} = (x_t - x_{t-1}, y_t - y_{t-1})$ and $V_t = (x_{t+1} - x_t, y_{t+1} - y_t)$ in consecutive pictures, namely, they must be greater than the threshold before they can be defined as an attack.

The total velocity of all objects must be summed if there are a total of n objects in a frame, and the following results are judged as candidate attack behaviors, namely,

Figure 2. The ROI objects of the first nine frames.

$$\sum_{i=1}^n V_{t-1} > T_{t-1} \text{ and } \sum_{i=1}^n V_t > T_t, \quad (15)$$

where T_{t-1} and T_t are thresholds.

In this study, a rectangular line was used to extract the body contour of the ROI after removing the background. The center point of the neck was set as the center point of the upper edge of the body contour. When it exceeds the outline range, it will be deleted, and then only the outline of the head and arms can be kept. As shown in Fig. 3, the vector V_a from the arm to the neck then can be obtained by calculating the center point of the arm contour and the vector V_c can be obtained by calculating the distance between the center of the body and the neck. Next, as shown in Fig. 4, the angle between the two vectors can be obtained as $\theta = \tan^{-1} \frac{V_a}{V_c}$. The detected behavior is regarded as an attack if the angle θ is greater than the threshold T_θ as shown in the following equation.

$$\theta = \sum_{i=1}^n \tan^{-1} \frac{V_{ai}}{V_{ci}} > T_\theta \quad (16)$$

where i is the number of ROI. In the end, the calculation of the minimized loss function uses cross-entropy in this paper.

$$L = -\sum_i (TPR_i + FPR_i) + \ln(PPV_i + NPV_i) \quad (17)$$

The optimal value is obtained by minimizing cross-entropy.

The thresholds T_{t-1} and T_t are then adaptively adjusted. The thresholds T_{t-1} and T_t are always adjusted if the accuracy obtained can be improved unless the accuracy value cannot be improved. Note that the accuracy is effectively continuously improved by adaptive adjustment of the thresholds of T_{t-1} and T_t . In other words, Eq. (15) only detects candidate attack behaviors, while Eq. (13) records the accuracy of the adaptively adjusted thresholds T_{t-1} and T_t . In addition, other data can be calculated, including false discovery rate (FDR), positive predictive value (PRV), predictive value (NPV), false missing rate (FOR), false-positive rate (FPR), negative true positive rate (TPR), true negative rate (TNR) and false-negative rate (FNR). These data can be passed through the following algorithms $\frac{FP}{TP+FP}$, $\frac{TP}{TP+FP}$, $\frac{TN}{FN+TN}$, $\frac{FN}{FN+TN}$, $\frac{FP}{FP+FN}$, $\frac{TP}{TP+FN}$, $\frac{TN}{FP+TN}$ and $\frac{FN}{TP+FN}$ are calculated, respectively. In these data, the most significant are the values of NPV and PRV. This is because NPV represents non-aggressive behavior and the prediction is also non-aggressive behavior, and PRV represents aggressive behavior and the prediction is also aggressive behavior. As shown in the flowchart in Fig. 5. The following steps are from the proposed system to determine if an attack has occurred.

Figure 3. The vector V_a is from the arm to the neck and the vector V_c between the center of the body and the neck.

Figure 4. The angle between vector V_a and V_c .

1. For input clip V , V is split into frames.
2. Remove the background by executing Eq. (2).
3. Obtain the ROI object using the saliency map and execute Eq. (12).
4. Execute Eqs. (15) and (16) to calculate the velocity and angle of the ROI object. If Eqs. (15) and (16) are true, go to step 6. If not, go to the next step.
5. Execute Eq. (17) to improve detection accuracy. If Eq. (17) continues to improve, the values of T_{t-1} and T_t in Eq. (15) and T_θ in Eq. (16) are adaptively adjusted. In another way, go to step 7.
6. Confirm that the behavior is offensive.
7. Confirm that the behavior is not offensive.

Experimental results

The algorithm proposed in this section performs simulations on several collected video sequences to evaluate the accuracy of judging aggressive behaviors, such as "Korea-students", "US-students", "Taiwan-students-Part I" and "Taiwan-students-Part II", their frame numbers are 50, 45, 50 and 50, respectively. A confusion matrix is used as an objective measure of accurate quality. Tables 1, 2, 3 and 4 are the comparison between the predicted results and the real conditions in the test clips of the proposed method. The proposed algorithm shows excellent accuracy values as can be seen from Tables 1, 2, 3 and 4. It can be seen that this method has obtained very good results on Accuracy. For Korea-students, US-students, Taiwan-students-Part I, and Taiwan-students-Part II films, it can be increased to 0.96, 0.98, 1, and 1, respectively. For the "US-students" and "Taiwan-students" sequences, the best accuracy of the proposed algorithm is 0.98 and the total average accuracy of all videos obtained by the proposed algorithm is 0.975. To further understand, compared with the best accuracy of the "US-students" and "Taiwan-students" sequences, the proposed algorithm reduces the minimum accuracy of the "Korea-students" sequence by only 0.04. In short, the proposed method has good performance and In short, the proposed method has good performance and only slightly reduces the performance for the "Korea-students" sequence. The comparison of NPV and PPV regarding the real conditions and prediction results of different videos in the proposed scheme is shown in Table 5, where the terms NPV and PPV represent the negative predictive value and the positive predictive value, respectively. As the terms PRV and NPV of the estimation accuracy of aggressive and non-aggressive behaviors, the proposed algorithm has the highest estimation accuracy in the sequence of "US-students" and

Figure 5. The flowchart of the proposed algorithms.

<i>Korea-students</i>	Total frames = 50	True condition	
		Aggressive	Non-aggressive
Predicted condition	Aggressive	22	2
	Non-aggressive	2	24
Accuracy		0.96	

Table 1. Comparison of the predicted results and real conditions of the proposed scheme with the test fragments of *Korea-students*.

<i>US-students</i>	Total frames = 45	True condition	
		Aggressive	Non-aggressive
Predicted condition	Aggressive	38	0
	Non-aggressive	1	6
Accuracy		0.98	

Table 2. Comparison of the predicted results and real conditions of the proposed scheme with the test fragments of *US-students*.

<i>Taiwan-students-Part I</i>	Total frames = 50	True condition	
		Aggressive	Non-aggressive
Predicted condition	Aggressive	38	1
	Non-aggressive	0	11
Accuracy		0.98	

Table 3. Comparison of the predicted results and real conditions of the proposed scheme with the test fragments of *Taiwan-students-Part I*.

<i>Taiwan-students-Part II</i>	Total frames = 50	True condition	
		Aggressive	Non-aggressive
Predicted condition	Aggressive	19	1
	Non-aggressive	0	30
Accuracy		0.98	

Table 4. Comparison of the predicted results and real conditions of the proposed scheme with the test fragments of *Taiwan-students-Part II*.

	Video	True condition	
		PPV	NPV
Predicted condition	<i>Korea-students</i>	0.92	0.92
	<i>US-students</i>	1	0.86
	<i>Taiwan-students-Part I</i>	0.97	1
	<i>Taiwan-students-Part II</i>	0.95	1

Table 5. The prediction results and real conditions of the proposed scheme with different videos are compared concerning PPV and NPV.

"*Taiwan-students*", which can be seen from Table 5. The NPV value of the "*US-students*" sequence is as high as 0.86 even though it is lower than other clips.

It can be seen from Table 6 that our proposed method proves that the objective performance evaluation in terms of accuracy is superior to Patwardhan²⁷, Veenendaal²⁸, and Goyal's²⁹ scheme, the accuracy is about 0.02–0.19. The proposed method outperforms the schemes of Patwardhan, Veenendaal, and Goyal with an

Video	Accuracy			
	Patwardhan	Veenendaal	Goyal	Proposed
<i>Korea-students</i>	0.91	0.78	0.94	0.96
<i>US-students</i>	0.93	0.81	0.96	0.98
<i>Taiwan-students-Part I</i>	0.92	0.79	0.95	0.98
<i>Taiwan-students-Part II</i>	0.94	0.82	0.96	0.98

Table 6. Compare the proposed scheme with Patwardhan, Veenendaal, and Goyal using accuracy videos for *Korean-students*, *US-students*, and *Taiwan-students*.

accuracy rate of about 0.02–0.18 dB in the *Korea-students* sequence. The proposed method outperforms the schemes of Patwardhan, Veenendaal, and Goyal with an accuracy of about 0.02–0.17 dB in the same comparison in the *US-students* sequence. The proposed method is better than the schemes of Patwardhan, Veenendaal, and Goyal. The accuracy rate of about 0.03–0.19 dB is also being compared in the sequence of Taiwanese students. The proposed method is better than the schemes of Patwardhan, Veenendaal, and Goyal with an accuracy rate of about 0.03–0.19 dB, and is also being compared in the sequence of *Taiwan-students*.

Conclusions

In this study, a motion recognition system using saliency map technology and background removal is proposed, and the accuracy is improved by confusion matrix and minimized cross-entropy. The ROI object of the frame is obtained by the method of saliency map and background removal. Whether it is an aggressive behavior is determined by detecting the angle between the velocity of the arm relative to the neck and the velocity of the center of the body relative to the neck. The accuracy of the proposed algorithm is improved by implementing a method based on a confusion matrix and minimizing cross-entropy. The attack behavior of the collected clips can be accurately detected and verified by the system based on the experimental results. It is proved by simulation that excellent accuracy performance can be achieved, such as several fragments of *Korea-students*, *US-students*, *Taiwan-students-Part I*, and *Taiwan-students-Part II*.

Received: 21 October 2021; Accepted: 8 February 2022

Published online: 23 February 2022

References

- Casas, J. F. *et al.* Early parenting and children's relational and physical aggression in the preschool and home contexts. *J. Appl. Dev. Psychol.* **27**(3), 209–227 (2006).
- Kupersmidt, J. B., Bryant, D. & Willoughby, M. T. Prevalence of aggressive behaviors among preschoolers in Head Start and community child care programs. *Behav. Disord.* **26**(1), 42–52 (2000).
- Landy, S. & Ray D. P. Toward an understanding of a developmental paradigm for aggressive conduct problems during the preschool years (1992).
- Tremblay, R. E. *et al.* The search for the age of 'onset' of physical aggression: Rousseau and Bandura revisited. *Crim. Behav. Mental Health* **9**(1), 8–23 (1999).
- Kokko, K. & Lea, P. Aggression in childhood and long-term unemployment in adulthood: A cycle of maladaptation and some protective factors 32a (2000).
- Kim, H. & Incheol, K. Dynamic arm gesture recognition using spherical angle features and hidden markov models. *Adv. Human Comput. Interact.* **20**, 15 (2015).
- Yao, J., & Jeremy R. Cooperstock. "Arm gesture detection in a classroom environment." In *Sixth IEEE Workshop on Applications of Computer Vision, 2002 (WACV 2002). Proceedings.* IEEE, 2002.
- Kim, D. *et al.* Vision-based arm gesture recognition for a long-range human–robot interaction. *J. Supercomput.* **65**(1), 336–352 (2013).
- Simão, M. A., Neto, P. & Gibaru, O. Unsupervised gesture segmentation by motion detection of a real-time data stream. *IEEE Trans. Ind. Inf.* **13**(2), 473–481 (2016).
- Lupinetti, K. *et al.* 3D dynamic hand gestures recognition using the Leap Motion sensor and convolutional neural networks. In *International Conference on Augmented Reality, Virtual Reality and Computer Graphics* (Springer, Cham, 2020).
- Akl, A., & Shahrokh, V. Accelerometer-based gesture recognition via dynamic-time warping, affinity propagation, and compressive sensing. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 2010.
- Akl, A., Feng, C. & Valaee, S. A novel accelerometer-based gesture recognition system. *IEEE Trans. Signal Process.* **59**(12), 6197–6205 (2011).
- Niezen, G., & Gerhard, P. H. Evaluating and optimising accelerometer-based gesture recognition techniques for mobile devices. *AFRICON 2009.* IEEE, 2009.
- Tsinganos, P., *et al.* A hilbert curve based representation of semg signals for the problem of gesture recognition. In *26th International Conference on Systems, Signals and Image Processing (IWSSIP).* 2019.
- Yeh, C.-H. *et al.* Vision-based virtual control mechanism via hand gesture recognition. *J. Comput.* **21**(2), 55–66 (2010).
- Chai, D., & Abdesselam, B. A Bayesian approach to skin color classification in YCbCr color space. In *2000 TENCON Proceedings. Intelligent Systems and Technologies for the New Millennium (Cat. No. 00CH37119).* Vol. 2. IEEE, 2000.
- Chen, S.-M., Chia-Hung, Y., & Ming-Te, W. Spatial-based video transcoding via visual attention model analysis. In *Conference on Computer Vision, Graphics, and Image Processing.* 2008.
- Ma, Y.-F., & Hong-Jiang, Z. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the Eleventh ACM International Conference on Multimedia.* 2003.
- Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **27**(8), 861–874 (2006).
- Powers, D. M. W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. -arXiv: [2010.16061](https://arxiv.org/abs/2010.16061) (arXiv preprint) (2020).

21. Stehman, S. V. Selecting and interpreting measures of thematic classification accuracy. *Remote Sens. Environ.* **62**(1), 77–89 (1997).
22. Qi, L. *et al.* Privacy-aware data fusion and prediction with spatial-temporal context for smart city industrial environment. *IEEE Trans. Ind. Inform.* **17**(6), 4159–4167 (2020).
23. Liu, Y. *et al.* A long short-term memory-based model for greenhouse climate prediction. *Int. J. Intell. Syst.* **37**(1), 135–151 (2021).
24. Liu, Y. *et al.* An attention-based category-aware GRU model for the next POI recommendation. *Int. J. Intell. Syst.* **20**, 20 (2021).
25. Qi, L. *et al.* Privacy-aware cross-platform service recommendation based on enhanced locality-sensitive hashing. *IEEE Trans. Netw. Sci. Eng.* **20**, 20 (2020).
26. Wang, F. *et al.* Robust collaborative filtering recommendation with user-item-trust records. *IEEE Trans. Comput. Soc. Syst.* **20**, 20 (2021).
27. Patwardhan, A., & Gerald, K. Aggressive actions and anger detection from multiple modalities using Kinect. [arXiv:1607.01076](https://arxiv.org/abs/1607.01076) (arXiv preprint) (2016).
28. Veenendaal, A. *et al.* Fight and aggression recognition using depth and motion data. *Comput. Sci. Emerg. Res. J.* **4**, 25 (2016).
29. Goyal, A. *et al.* Automatic border surveillance using machine learning in remote video surveillance systems. In *Emerging Trends in Electrical, Communications, and Information Technologies* 751–760 (Springer, 2020).

Author contributions

M.-T.W. wrote the main manuscript text.

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.-T.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022