Sensors and Materials, Vol. 36, No. # (2024) 1–17
MYU Tokyo

S & M ####

Healthcare System from Multisensor Collaboration and Human Action Recognition

Hongwei Gao,[1,2*] Xuna Wang,[1**] Zide Liu,[1] and Yueqiu Jiang[3]

[1]School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110159, China
[2]China State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China
[3]School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China

(Received November 21, 2023; accepted January 5, 2024)

Keywords: IoT, action recognition, pose estimation, remote healthcare

Over the past few decades, wearable sensor technology has played a pivotal role in patient information acquisition. A new paradigm for unconstrained medical data collection emerged with noncontact sensors such as Kinect and industrial-grade RGB cameras. These innovations have enhanced patient experiences and offer vast potential for long-term monitoring, telemedicine, and remote healthcare in underserved areas. However, while many research efforts have zeroed in on specific components, such as sensor development or algorithm design, this has occasionally led to noncontact systems' reliability, accuracy, and connectivity challenges. Consequently, there is a pressing need for research that adopts a more holistic approach, ensuring the optimal integration of sensors and algorithms. In this study, we introduce a method of constructing a noncontact diagnostic system powered by deep learning vision algorithms that showcase strong resilience against viewpoint changes and obstructions in human motion identification and assessment. Using four RGB cameras, we capture human dynamics and leverage a pose estimator to generate comprehensive 3D human postures. Afterward, these postures are refined to aid in subsequent behavior prediction. Our multitask-trained model significantly bolsters the system's adaptability to posture discrepancies. Notably, this noncontact diagnostic system thrives in challenging environments, such as 360-degree surveillance, intricate situations, and low-light settings where traditional sensors often falter. In addition, we have assembled a multiview autism behavior dataset. Through it, our embedded deep learning algorithm showcases exemplary action category recognition (reaching up to 95.09%), highlighting further its practical implications.

1. Introduction

In healthcare, wearable devices such as electromyogram sensors generate detailed physiological data and behavioral insights.(1) These devices, while helpful, often cause discomfort to patients and restrict their movement, hindering the collection of everyday data. To address these challenges, the continuous development of artificial intelligence and its integration

*Corresponding author: e-mail: ghw1978@sohu.com
**Corresponding author: e-mail: xuna_emmm666@163.com

ISSN 0914-4935

with visual sensing and behavior recognition technology has emerged as a promising approach for future healthcare.[2] However, the application of this technology faces obstacles, particularly in managing the unpredictable behaviors of patients in open spaces.

Although IoT and AI are independent technological domains, they are increasingly linked in many cases. This convergence enhances the accuracy of behavior recognition technology, which benefits from data collected from real-time multisensory sources. Furthermore, data exchange through IoT devices can bring more efficient solutions to the healthcare sector. Building upon these advancements, in this study, we explore an advanced method that combines behavior recognition technology with IoT.[3,4] We aim to maximize the potential of visual sensors, thereby significantly enriching and enhancing behavior monitoring and treatment plans.

Nevertheless, given its interdisciplinary intricacies, much research overlooks the synchronization between sensor configurations and AI paradigms. Medical or sensor researchers gravitate toward conventional deep-learning frameworks when crafting algorithms. For instance, Cook *et al.*[5] unveiled an autism behavior surveillance rooted in skeletons. They exploited a random forest methodology for behavior category prediction, clocking in an accuracy rate of merely 70%. Conversely, Kongjun and Yaoxi[6] discussed a medical monitoring technique hinged on grayscale value segmentation for human object identification without delving deeper into motion evaluation or identification. Although behavior recognition is advancing within the computer vision sector, many models cater to diverse challenges. However, a recurring issue is these models' validation primarily on public datasets rather than real-world scenarios.[4,7,8]

It is crucial to prioritize research on the concurrent design of AI algorithms and IoT infrastructures, especially in behavior-focused healthcare frameworks. These frameworks can typically be categorized into behavior detection and evaluation.[9] The skeleton-based method is widely used for assessing behavior and identifying actions, as it helps minimize background interference in difficult-to-reach settings. However, the contactless vision technology's pose estimation accuracy is not as high as that of wearable devices, posing challenges to deploying existing models trained on actual pose data in practical applications. Moreover, vision sensors in unstructured medical environments often capture incomplete or challenging perspective objects, significantly impairing the model's ability to comprehend motion.[10,11]

On the basis of this, we introduce a methodology for a noncontact diagnostic system that exploits deep-learning visual algorithms. As depicted in Fig. 1, the system employs cameras to collect behavioral information from multiple angles. This information is then processed by behavior analysis technology to obtain 3D posture coordinates and behavioral categories that are useful for medical assessment. Subsequently, these analyzed results are transmitted to a terminal for further operation within the relevant medical application. During the system's development phase, Kinect is used to collect comprehensive 3D posture data, which guides the training of the behavior analysis model to enhance its robustness in varied application environments and filming angles. Additionally, the algorithm design efficiently utilizes data from multiple sensors and achieves simultaneous behavior detection and assessment. By applying our designed system, we have compiled a multiangle dataset on autistic behaviors. Experiments demonstrate that our algorithm outperforms traditional and state-of-the-art technologies, achieving outstanding recall and precision metrics results. Moreover, by balancing detection accuracy and system cost, we

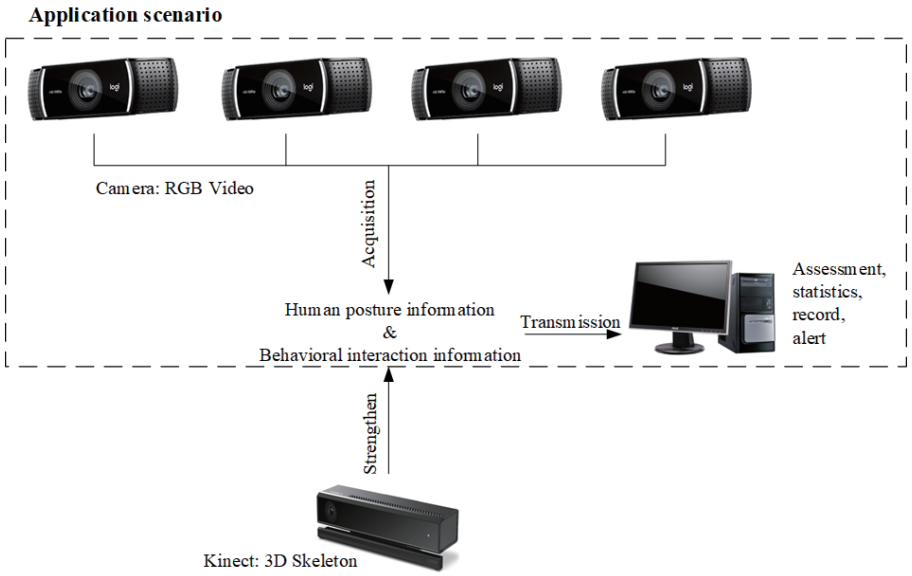

Fig. 1. Health behavior monitoring system based on visual sensing.

determined that the optimal arrangement of four visual sensors was the most effective configuration.

This work has contributed the following to healthcare systems:

1. A visionary noncontact system. Quadruple visual sensors document video behavioral sequences, while behavior analysis algorithms deliver premium viewpoint 3D landmarks and behavior classifications, ensuring robustness against coordinate offset and visual obstruction. The acquired behavioral data are channeled for assessments or analytics, offering real-time feedback to healthcare experts or caregivers
2. An evolved behavior analysis algorithm tailored for diagnostic systems. Initially, the pose estimation algorithm, enriched with a 2D-to-3D transition module, furnishes more precise 3D poses by capitalizing on data harvested from multiple viewpoints. Then, a viewpoint adjustment component has been employed to lay the groundwork for efficient behavior prognostication. Finally, the simultaneous training of 3D pose approximation and behavior forecasting fortifies the model's resilience to pose coordinate fluctuations.
3. A proprietary dataset and empirical trials. We gather video footage spanning five perspectives for autism behavior, accompanied by a 3D skeletal sequence from a singular angle. Drawing upon this dataset, we attest from our experiments that the setup with four visual sensors is the most productive, registering an average detection precision of 95.09%, which eclipses state-of-the-art methodologies.

The rest of the paper is organized as follows. In Sect. 2, we briefly introduce vision sensing systems in medical applications, with a special focus on behavior recognition in healthcare. Then, we describe the main methods of the proposed model in Sect. 3. The main content includes two parts: the adaptive behavior analysis algorithm and the visionary noncontact diagnostic

system. To verify that the model can be applied to the healthcare system, we conducted two sets of experiments in Sect. 4, testing both the algorithm and its application performance. Finally, we summarize the model and its development space.

2. Related Works

To introduce our system and its context, we first discussed the widespread application of computer vision in the medical field; this section further delves into the research background of our paper, emphasizing the specific application of behavior recognition technology in healthcare.

2.1 Vision sensing systems in medical applications

Visual sensing systems have become indispensable tools in the medical sector, offering capabilities spanning diagnosis, treatment, surveillance, and patient welfare enhancement.[12] A closer look at the diverse healthcare applications of visual sensing systems reveals the following:

Pathological Image Analysis: Visual sensing systems can visualize organs and tissues for diagnosis and disease tracking. To extract vital characteristics of multiple sclerosis, Silvia *et al.*[13] employed noninvasive visual evoked potential and optical coherence tomography. This study further supports the noninvasive exploration of potential treatment methods, which can promote remyelination in demyelinating diseases.

Remote Patient Monitoring: Visual sensing systems allow remote diagnosis and record the patient's medical information. Malasinghe *et al.*[14] harnessed a camera to document physiological reactions stemming from heartbeats. This led to the actualization of a video-centric remote heart rate determination method.

Vision Systems for Robots: Medical robots integrate mechanics, electronics, and computing to perform various healthcare tasks.[15] Bahar *et al.*[16] embedded an algorithm rooted in metric learning and dynamic time warping into a robotic entity. This system autonomously identifies and evaluates pivotal joints, delivering feedback based on spatial–temporal disparities between pediatric patients and established models.

Virtual reality and augmented reality (AR): These techniques can bring participants into different teaching situations and promote innovation in teaching forms. Ahlers *et al.*[17] proposed an AR computer interaction system based on voice and gesture interaction to improve children's cognitive abilities. Tsai *et al.*[18] used role-playing games from a third-person perspective to teach social skills and help deepen understanding of basic emotions.

To summarize, vision sensors are widely used in the medical field owing to their outstanding performance. Moreover, the resulting noninvasive medical paradigm significantly improves the patient experience. Therefore, vision-sensor-based medical systems have vast potential for development.

2.2 Behavior recognition in healthcare

Human actions inherently bear the hallmark of nonverbal cues that, when decoded, can elucidate an individual's physiological and psychological facets. These movements, stemming

from limbs, hands, and myriad other body components, are brimming with invaluable data crucial in healthcare.[19] Venturing into the domain of healthcare robotics, key avenues of exploration span geriatric care, anomalous behavior surveillance, kinetic rehabilitation, and cognitive therapy.

Geriatric Medicine: A primary concern underpinning senior care is the detection of falls, given the perils they pose to older people living autonomously. Sree and Jeyakumar[20] made notable endeavors in this direction and introduced fall detection leveraging background subtraction and finite state machines pivoted on metrics derived from bounding boxes. Further enriching this narrative, Yadav *et al.*[21] launched a system wherein video-extracted skeletal coordinates were channeled through convolutional neural networks and gated recurrent units sequentially, aiming to oversee daily activities and detect falls.

Abnormal Behavior Monitoring: Autism spectrum disorder (ASD) is characterized as semi-voluntary recurrent motions. These actions in unbridled environments have been studied by Negin *et al.*,[22] who deployed long short-term memory (LSTM) networks to discern skeletal sequence progression over time. For the same task, Liu *et al.*[23] proposed a dual-channel multiscale depth-wise separable convolutional neural mechanism designed to distill intricate spatial features from abnormal gaits while ensuring computational economy.

Sports Rehabilitation: In athletic recovery, intricate actions and their nuances become pivotal for diagnosis and treatment. Zunino *et al.*[24] leverage an LSTM-enhanced GoogleNet to identify ASD patients based on the action of grasping and placing a bottle. Inspired by this work, Sun *et al.*[25] accentuated spatial information extraction using spatial attention bilinear pooling, hitting an accuracy pinnacle of 82.56%. Targeting patients with chronic obstructive pulmonary disease, Fan *et al.*[26] proposed a six-action respiratory training regime and deployed a hybrid convolutional neural network-LSTM framework to analyze recorded kinetic data.

Cognitive Rehabilitation: The heart of mental recovery revolves around gauging the degree of engagement, especially among pediatric patients.[27] To this end, Anzalone *et al.*[28] spotlighted parameters describing children's motions, encompassing facets such as head and body kinetics, gaze metrics, and kinetic energy. Enhancing the observation process, Shi *et al.*[29] delved into clinical settings where children intermittently handed over toys during interactional games. They designed the OstAD network to combine contextual features and derive local frame-level insights.

When harnessed and interpreted correctly, the intricate information of human motion can be an invaluable asset in the healthcare domain. Behavior analysis algorithms are designed for different medical scenarios in the above-related work. Most algorithms obtain behavior information based on familiar pose estimation or behavior recognition models and then calculate the behavior characteristics for medical diagnosis according to the application scenarios. Therefore, a general action understanding can be built, and corresponding behavior feature calculation units can be embedded according to different medical procedures.

3. Materials and Methods

As shown in Fig. 2, this methodology encompasses three pivotal elements: sensor arrangement, behavior analysis algorithms, and healthcare data manipulation. Visual sensors are

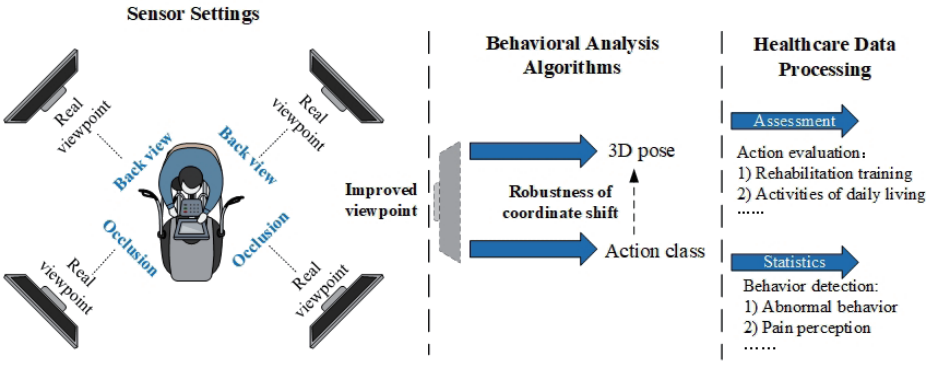

Fig. 2. (Color online) Healthcare system from multisensor collaboration and human action recognition.

tactically positioned in the environment to record video data from various angles. The behavior analysis algorithms augment the 3D pose coordinates, optimizing for better posture viewpoints and subsequently facilitating the categorization of behaviors based on the gathered intel. Following this, the 3D pose is employed for evaluating actions, such as rehabilitation exercises or daily routines. Concurrently, the identified behavior categories are harnessed for statistical analyses, covering aspects such as aberrant behavior or pain discernment.

3.1 Visionary noncontact diagnostic system

The noncontact visual system outlined in this study presents a nuanced approach to addressing inherent challenges in unstructured healthcare environments. Its central aim revolves around offering reliable behavior assessment and statistical analysis across diverse settings. A closer examination elucidates the following:

Hardware Configuration: Figure 2 shows that the system's hardware foundation comprises four strategically placed visual sensors. Positioned at each corner of the designated environment, these sensors ensure a comprehensive coverage of the entire area. Patients' movements are unrestricted against a multifaceted background in spaces devoid of constraints. Consequently, capturing varied postural perspectives is complex, often compounded by concerns such as occlusion and suboptimal lighting conditions. In the subsequent sections, we will delve deeper into the intricacies of the formulated behavioral analysis algorithms.

Algorithmic Workflow: Figure 3 offers a schematic representation of the algorithm's operational flow. Initially, the system processes incoming data to pinpoint 3D pose key points, essential for effective behavior recognition. Utilizing these 3D poses, the system discerns the behavioral category, subsequently ascertaining its significance (i.e., its importance for detection). For behaviors deemed valuable, a subsequent determination determines the need for detection versus assessment. Detection-driven behaviors witness a meticulous gathering of information, focusing on attributes such as frequency, magnitude, and duration. This compiled data, coupled with behavioral insights, is relayed to remote control apparatuses. For behaviors earmarked for assessment, a rigorous evaluation quantifies the behavior's adherence to standards, leveraging

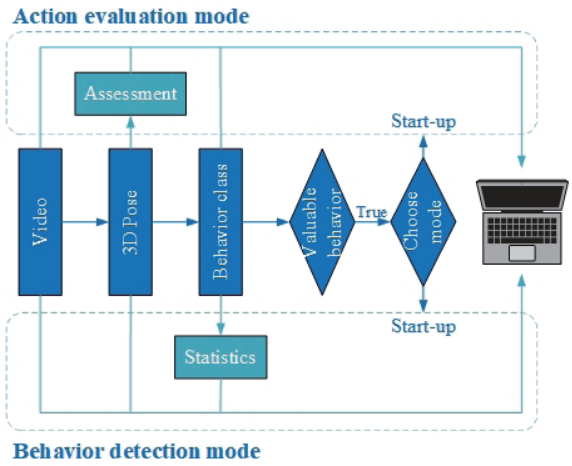

Fig. 3. (Color online) Algorithm workflow.

the 3D pose data. These assessments, alongside relevant behavioral data, are then dispatched to remote devices. A distinctive system feature pertains to its alert mechanism. The occurrence of specified risk-associated behaviors triggers automatic notifications to concerned stakeholders or doctors, caregivers, or family. The implementation of this system mainly includes the following details:

1. Data Storage Strategy: Addressing the challenge of securely archiving extended durations of patient video logs, the system employs a dual-pronged strategy. A primary compression via the H.264 video codec ensures substantial size reduction without compromising video clarity. Videos encapsulating targeted behaviors benefit from high-speed storage solutions, while the remainder transitions to cost-efficient long-term storage options. Specifically, the data inconsistent with the target behavior will be migrated to offline or tape storage mode after two days to optimize storage spend.
2. System Stability & Availability: The system prioritizes session consistency, integrating an Internet Protocol (IP) hash-centric load balancing technique.[30] The hashing process involves using the Cyclic Redundancy Check 32 algorithm to convert IP addresses into a numerical hash value. Then, by dividing the hash value by the number of servers and using the remainder, a server is selected from the pool of available servers. Finally, the traffic is forwarded to this server. This mechanism hinges on hashing client IP specifics, guaranteeing that successive requests from an identical client invariably route to the same server.
3. Scalability: Catering to medical establishments of varying magnitudes and fluctuating patient volumes, the system exploits a microservices architectural paradigm. This divides the system into discrete, autonomous services, facilitating scalable adjustments to individual service instances without overarching system disruptions. Each microservice embodies a distinct functional role spanning video data preprocessing, posture and behavior analyses, data analytics, communication interfaces, user-centric interfaces, and beyond. Microservices are containerized using Docker technology, encapsulating them in a consistent environment.

They locate each other for communication through server-side discovery, where each service registers its location in a service registry, and clients query this registry to find services. Microservices communicate with each other using restful application programming interfaces.

4. Data Privacy & Security: Patient confidentiality remains paramount. To this end, the system incorporates the transport layer security protocol, ensuring that video data transmissions remain encrypted and safeguarded against potential breaches or manipulations during transit. In addition, encryption measures extend to storage mediums hosting nontarget behavior data, precluding inadvertent data disclosures about daily patient activities.

3.2 Adaptive behavior analysis algorithm

3.2.1 Algorithm architecture

The algorithm's specific workflow is illustrated in Fig. 4. Videos from different perspectives use a 2D pose estimator to generate 2D keypoint coordinates and confidences. The keypoint information from all views is input to the 3D pose estimator, which, along with keypoint confidence features, produces accurate and complete 3D pose coordinates. The 3D pose is then transformed into a more recognizable viewpoint using the viewpoint transformation module, and the behavior predictor outputs the behavior category. The modules after the 2D pose estimator are jointly trained, utilizing the relatively low computational cost of ShiftGCN's spatiotemporal information extraction module.[31] The algorithm's structure is outlined in the lower half of Fig. 4 as follows.

2D pose estimation uses OpenPose,[32] resulting in 18 key point coordinates (x, y) and confidences. The 3D pose estimation inputs 2D poses from four perspectives, and through network learning, it achieves mapping from 2D perspectives to a 3D perspective. It fuses the mapping coordinates of the same joint based on different confidence levels. All the above processes are implemented through self-learning by the network. This module employs a U-shaped architecture.[33] Given the continuity of actions, consecutive frames exhibit high correlations in the pose information. The framework first reduces dimensionality along the temporal dimension to extract spatiotemporal information progressively, and then it up-samples the data to predict 3D coordinates. During the up-sampling process, skip connections supplement the lost information at the same dimensionality as the down-sampling part. The up-sampling rates for the temporal dimension are 2 and 1/2. This approach effectively utilizes long-term and short-term material information.

In the behavior prediction part, the viewpoint transformation module is initially employed to convert the viewpoint into one suitable for distinguishing behavior categories. Subsequently, a multilevel feature fusion module is used to extract behavior features. In each stage, the first unit sets the temporal dimension stride to 2, and correspondingly, the down-sampling is set to 1/2 in the feature fusion unit. The channel attention mechanism is used in the feature fusion unit to generate the fusion weights of different feature channels. It is implemented as follows. First, global average pooling obtains aggregated features for each channel. Then, one-dimensional

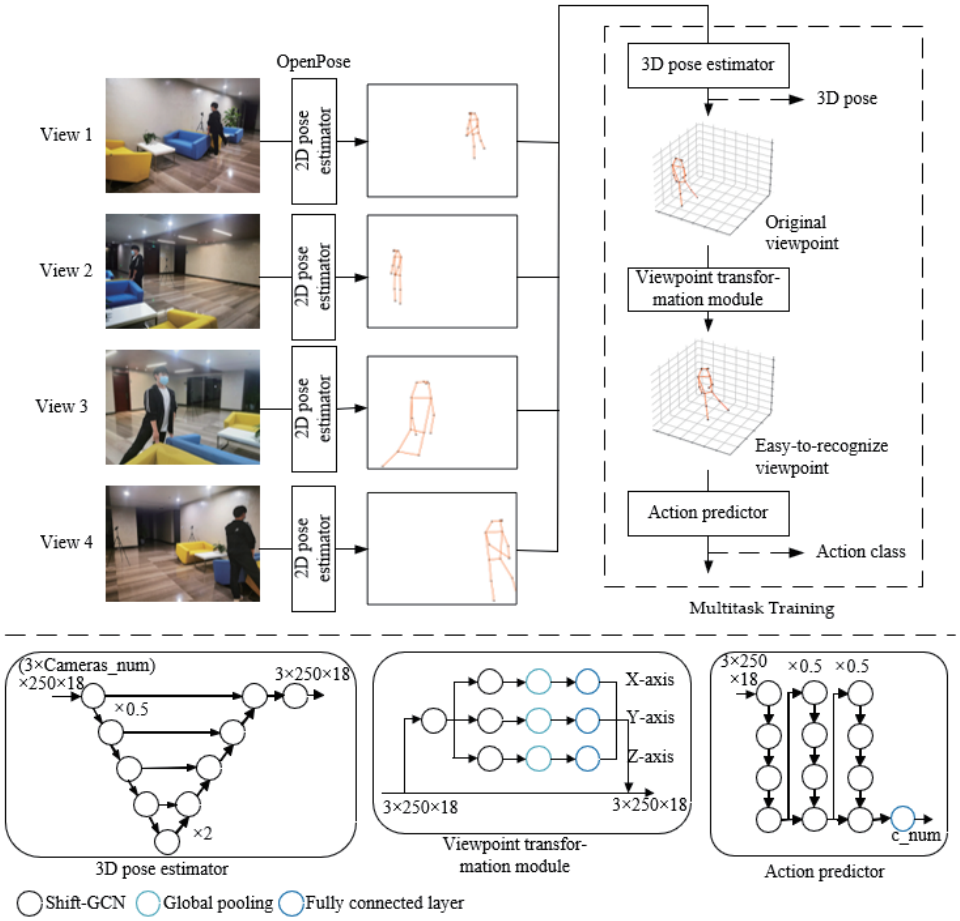

Fig. 4. (Color online) Algorithm workflow and module structure.

A-Shift Conv is used to obtain channel weights. Finally, the sigmoid function maps the weights to the range of 0–1. Finally, a fully connected layer with an output size equal to the number of categories is used to output the prediction probabilities for each category.

Pose estimation and behavior prediction are tasks related to understanding human movement, and they have a strong correlation, enabling mutual enhancement. Behaviors are classified on the basis of human posture coordinates and their changes, while the posture coordinates and changes are constrained by the types of behaviors. In our approach, the 2D pose estimation component does not participate in joint training. The training of the 3D pose estimator that follows will optimize the 2D coordinates. Since a well-initialized pose estimation component aids the convergence of the behavior prediction component, it is advisable to train the 3D pose estimation part separately for 30 epochs before jointly training it with the behavior prediction part. Although pose estimation and behavior prediction are related tasks, they have different weight distributions. Pose estimation focuses on local information, whereas behavior prediction focuses on global information. Therefore, we alternate the training of pose estimation and

behavior prediction, and during the training of the behavior prediction component, we freeze the lower layer weights of the 3D pose estimation module.

3.2.2 Viewpoint transformation module

The viewpoint transformation module, inspired by the work of Zhang *et al.*,[34] is a critical component in this study's algorithm. Unlike the original work, in this research, we focus on designing the module based on the type of coordinate system change and its effect on different coordinate dimensions. The central concept behind this module is to generate a set of parameters through network learning for each object; in the equation provided, r represents the rotation angle around a particular axis of the view coordinate system O and d represents the translation distance along a specific axis. The 3D joint coordinates in the new view coordinate system O' are represented as

$$\hat{V}'^{(:,m)} = R\left(\hat{V}'^{(:,m)} - D\right) \in \mathbb{R}^{N_t \times D_3 \times N}. \tag{1}$$

In this equation, N_t denotes the number of video frames. R is represented as

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(r_x) & -\sin(r_x) \\ 0 & \sin(r_x) & \cos(r_x) \end{bmatrix} \times \begin{bmatrix} \cos(r_y) & -\sin(r_y) & 0 \\ \sin(r_y) & \cos(r_y) & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \cos(r_z) & 0 & \sin(r_z) \\ 0 & 1 & 0 \\ -\sin(r_z) & 0 & \cos(r_z) \end{bmatrix}. \tag{2}$$

The parameter vector T, responsible for coordinate system changes, is designed following the rules governing how it affects the coordinates. If the coordinate system undergoes translation along a particular axis, it affects the position coordinates in that dimension. If the coordinate system rotates around a specific axis, it affects the function coordinates in the other two dimensions. Taking the z-axis as an example, changes in the z-axis can be achieved through translation along the z-axis and rotation around the x- or y-axis. Accordingly, the primary features extracted generate vectors separately for coordinate changes in the x, y, and z dimensions. For the z-coordinate dimension, the output transformation vector is $T_z = \{r_{x|z}, r_{y|z}, d_z\}$, where $r_{x|z}$, $r_{y|z}$ repres ents the rotation angles around the x- and y-axes owing to changes in z-coordinate dimension. The final rotation vector is denoted as $r_z = (r_{x|z} + r_{y|z})/2$.

The workflow of the viewpoint transformation module is illustrated in the lower half of Fig. 4. The input coordinate data structure is *(N × M, C, T, V)*. It goes through a feature extraction layer to obtain the overall coordinate features and input them into three parallel single-dimensional transformation deduction channels. First, the feature extraction layer provides features specific to the dimension of interest. Subsequently, a global pooling layer with an output size of 1×1 calculates the averages of the extracted features channelwise. Finally, a fully connected layer with three neurons outputs the transformation vector T_*.The three sets of transformation vectors are combined to generate the parameter vector T, which performs rotation and translation transformations on the coordinate system based on Eq. 1.

4. Results and Discussion

4.1 Dataset creation

The prevalence of ASD is on the rise worldwide and has rapidly developed into a global public health crisis. Children with long-term chronic conditions associated with ASD often experience varying degrees of social impairments. However, owing to significant delays in the early screening, diagnosis, and treatment of autism, the rate of early intervention is relatively low.[35] One prominent clinical feature of ASD is stereotyped behavior. Detecting atypical behavior in children can aid in the early identification of autism risk.

In this study, we collected a multiview dataset of autism behaviors through self-collection inspired by the Autism Spectrum Disorder Behavioral Dataset.[36] The dataset consists of four action classes: arm flapping, hand movements, head-banging, and spinning. Each action class includes 180 samples captured by three different subjects, with each issue providing 60 samples. The samples were randomly split into a training set, a validation set, and a test set in an 8:1:1 ratio. Each sample comprises video clips from five different viewpoints and 3D skeletal data from one perspective, all collected simultaneously. Each sample lasts approximately 10 s, with 25 skeletal sequences generated per second. As illustrated in Fig. 5, the videos were recorded using C920 PRO HD WEB CAMs, with each camera placed at one of the four corners of the scene. The 3D skeletal data and the frontal video screen were generated by Kinect V2, which was positioned at the front center of the scene. While performing actions, the subjects were required to stay within the orange frame area to ensure that Kinect recorded their entire bodies. Note that Kinect was used exclusively for training the 3D pose estimator in this research, and the proposed system does not include Kinect.

4.2 Model training

All our experiments were conducted on NVIDIA-SMI 516.95 with CUDA version 11.7. The models in the system were implemented using PyTorch 1.12.1 and Python 3.9. We trained our

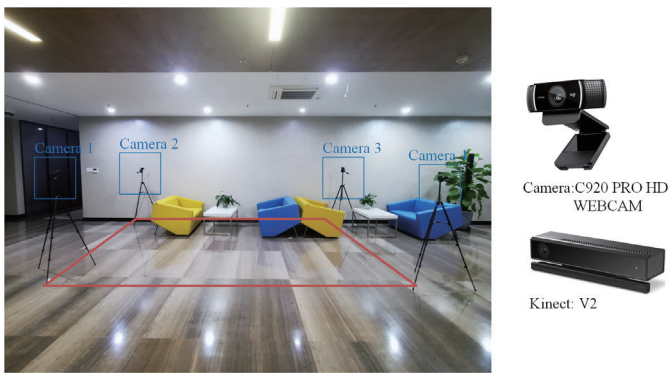

Fig. 5. (Color online) Acquisition equipment and setting (The application does not include Kinect).

models using the stochastic gradient descent (SGD) optimizer, with the training process divided into two stages, as follows.

3D pose estimation: We trained the 3D pose estimation model in this stage. We used Nesterov momentum with a rate of 0.9 and a weight decay of 0.0005. The initial learning rate was set to 0.01 and decayed by 0.1 at the 70th, 95th, and 100th iterations. A batch size of 16 was used, and we trained more than 105 epochs. The training for the behavior recognition part began after an initial pretraining of 30 epochs for 3D pose estimation.

Behavior recognition training: Behavior recognition training was performed alternatingly with the 3D pose estimation part. When training for behavior recognition, all layers of the 3D pose estimation part before the final layer were frozen. We used the SGD optimizer with Nesterov momentum set at 0.9 and weight decay at 0.0001. The initial learning rate for this stage was 0.1 and was decayed by 0.1 at the 35th, 55th, and 70th iterations. Similar to the 3D pose estimation part, a batch size of 16 was used, and the training process spanned 75 epochs.

4.3 Algorithm testing

We tested our model on the multiview autism behavior dataset and compared it with classical and state-of-the-art models. The videos generated by cameras 1–4 were input into each model. To ensure a fair comparison, single-view models combined the predicted probabilities from different views by averaging to obtain behavior categories.

As shown in Table 1, our model achieved the best average and class-specific accuracy, with an average accuracy exceeding VTN[37] (ranked second) by 3.87%. Except for the arm-swinging action, the prediction accuracy for all actions exceeded 90%, with a prediction accuracy of 94.12% for the head-banging action. Our model excelled in detecting the arm-swinging action compared with other activities, demonstrating superior performance even in multiview applications. The algorithm proposed in this research, which utilizes 2D poses from different viewpoints for 3D pose estimation, helps mitigate the impact of occluded joints on recognition accuracy.

We visualized the confusion matrices for each model. As shown in Fig. 6, our model's predictions are more concentrated on the diagonal, indicating that the model has the best recall. This means that when a patient exhibits the target behavior, the system has a higher probability of recognizing this category and providing feedback to the doctor. Additionally, our model's

Table 1
Prediction accuracy on the custom dataset.

Methods	Mean accuracy	Accuracy of each class			
		Arm flapping	Head banging	Spinning	Hand action
Resnet[38]	83.39	75.21	83.59	88.66	87.19
Swin-Trans[39]	85.13	76.44	85.51	90.21	89.58
3D-HAR[40]	88.35	79.93	87.46	94.45	92.01
Semi-CNN[41]	90.38	81.86	89.37	96.73	93.97
VTN[37]	91.22	84.42	91.77	96.79	93.24
Ours	95.09	89.37	95.24	99.12	Ours

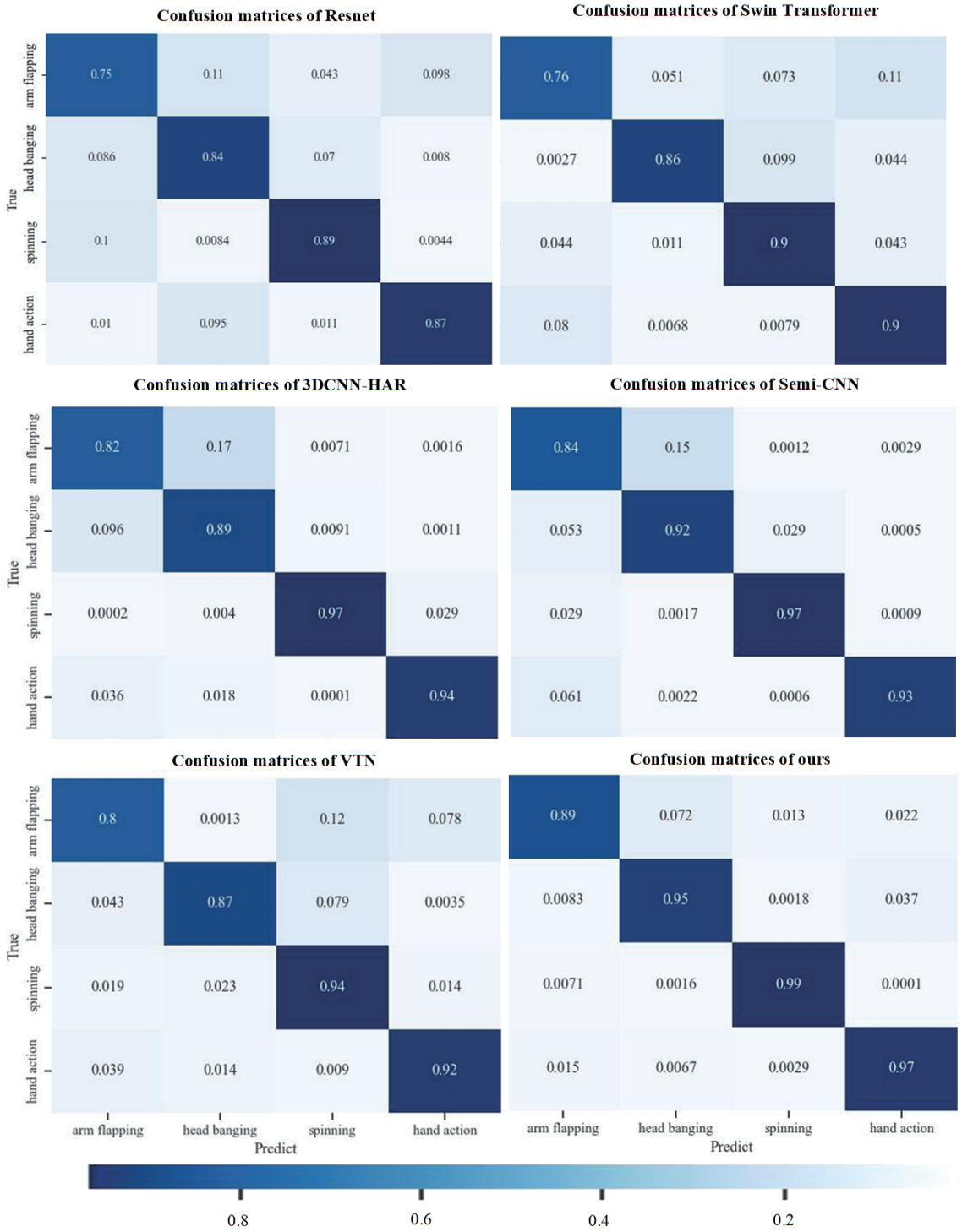

Fig. 6. (Color online) Confusion matrices.

precision is higher than those of other models, reducing the likelihood of false positives. The arm-swinging and head-banging actions can be easily confused since both activities involve swinging the head and both arms. However, our model's error rates are below 5%, indicating high performance.

4.4 Application testing

To assess the system's application capabilities, we conducted tests with four different sensor configurations, each ensuring complete coverage of the scene area. As shown in Fig. 7, we provide the model's sensor configurations and corresponding accuracy curves. We simulated different scenarios by feeding the model with videos from various camera setups. For instance, videos from cameras 1 and 3 were used for training and testing in a two-sensor system.

The results indicate that the system's performance gradually improves as the number of viewpoints increases. Notably, the model's performance is most significantly enhanced when the number of views changes from 2 to 3, with an improvement of 3%. However, when the number of viewpoints is 4 or 5, the model's improvement is relatively minor, around 0.2%. We set the number of cameras to 4 to reduce equipment costs and adopted the arrangement shown in Fig. 7.

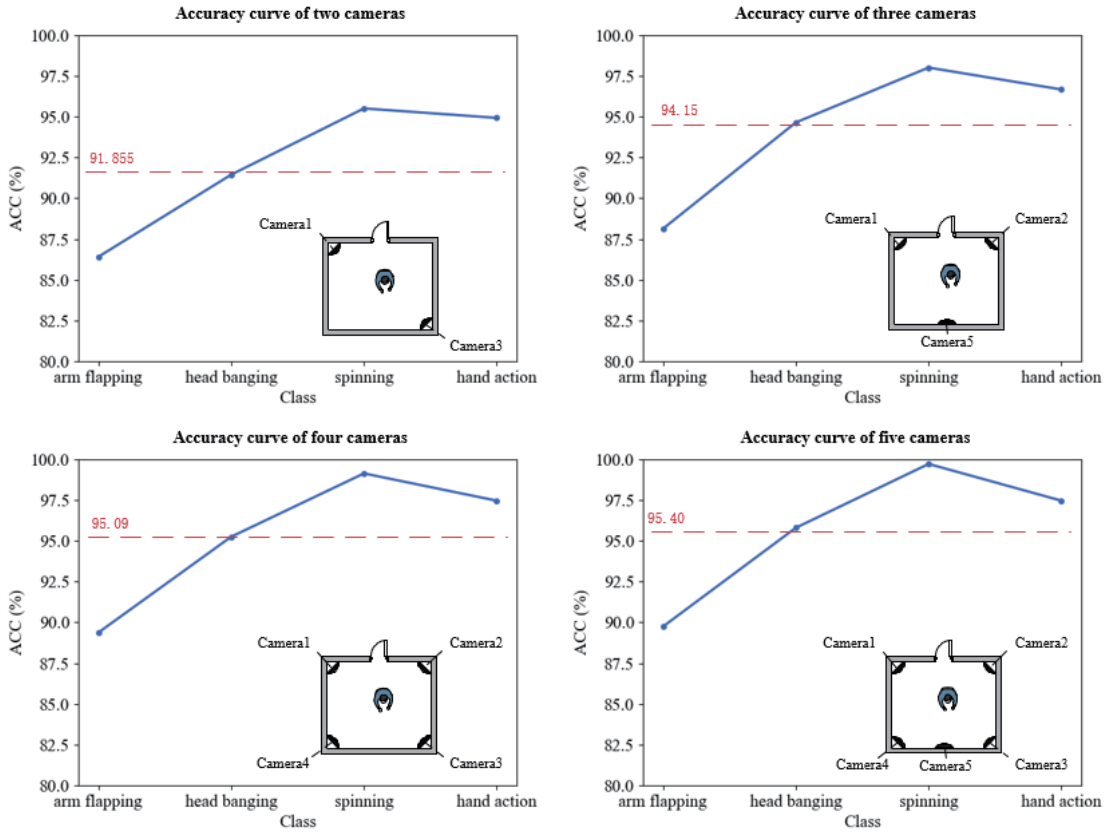

Fig. 7. (Color online) Accuracy curves for different hardware configurations.

5. Conclusions

We introduced a system for behavior detection and analysis in response to the challenges presented by nonstructured medical paradigms. This system primarily comprises four visual sensors, each placed at different viewpoints, which capture video data. This video data is then processed by behavior recognition algorithms tailored for integration into healthcare systems. Behavior recognition algorithms based on video data face three common challenges in healthcare applications. Our approach has been specifically designed to address these challenges. To address the issue of obscured body parts and joints, we employ multiview 2D pose estimation along with confidence scores to generate a comprehensive 3D pose. We implement a viewpoint transformation module to tackle the problem of behavior recognition becoming difficult owing to camera viewpoints. This module adjusts the generated 3D posture coordinates to a more easily recognizable view. To alleviate the issue of coordinate offsets in pose estimation algorithms, we employ the combined training of 3D pose estimation and behavior recognition. This combined training enhances the system's robustness against balances. We tested the system using a self-collected dataset of autism behavior. The results indicate that our method achieved an accuracy of 95.09%, surpassing current state-of-the-art algorithms. This underscores the effectiveness of our system in mitigating the challenges presented by complex environments, making it a valuable asset for medical applications.

Acknowledgments

This work was supported by the Liaoning Province Higher Education Innovative Talents Program, grant number LR2019058, Liaoning Province Joint Open Fund for Key Scientific and Technological Innovation Bases, grant number 2021-KF-12-05, the Zhejiang Provincial Natural Science Foundation of China, grant number LQ23F030001, and the Open Fund of State Key Laboratory of Robotics (Grant No. 2023-O03).

References

1 C. C. Chen: Interact J. Med. Res. **9** (2020) e19776. https://doi.org/10.2196/19776
2 H. Albert, M. Arnold, and F.-F. Li: Nature **585** (2020) 193. https://doi.org/10.1038/s41586-020-2669-y
3 S. Farahnaz, B. Ali, and S. Nasrin: J. Biomed. Inf. **103** (2020) 103383. https://doi.org/10.1016/j.jbi.2020.103383
4 J. Yu, H. Gao, Y. Chen, D. Zhou, J. Liu and Z. Ju: IEEE Trans. Hum.-Mach. Syst. **52** (2022) 784. https://doi.org/10.1109/THMS.2022.3144951
5 A. Cook, B. Mandal, D. Berry, and M. Johnson: Proc. 2019 IEEE Int. Conf. Data Science and Advanced Analytics (IEEE, 2019) 504–510. https://doi.org/10.1109/DSAA.2019.00065
6 B. Kongjun and B. Yaoxi: J. Healthcare Eng. **2021** (2021) 6549891. https://doi.org/10.1155/2021/6549891
7 J. Yu, H. Gao, D. Zhou, J. Liu, Q. Gao, and Z. Ju: IEEE Trans. Cybern. **52** (2022) 13738. https://doi.org/10.1109/TCYB.2021.3114031
8 A. Piergiovanni, W. Kuo, and A. Angelova: Proc. 2023 IEEE/CVF Conf. Computer Vision and Pattern Recognition (IEEE, 2023) 2214–2224. https://doi.org/10.1109/CVPR52729.2023.00220
9 Aledhari, R. Razzak, B. Qolomany, A. Al-Fuqaha, and F. Saeed: IEEE Access. **10** (2022) 31306. https://doi.org/10.1109/ACCESS.2022.3159235
10 J. Yu, H. Gao, Y. Chen, D. Zhou, J. Liu, and Z. Ju: IEEE Trans. Cognit. Dev. Syst. **14** (2022) 1654. https://doi.org/10.1109/TCDS.2021.3131253.

11 J. Yu, Y. Xu, H. Chen, and Z. Ju: IEEE Trans. Neural Networks Learn. Syst. (2022) 1–13. https://doi.org/10.1109/TNNLS.2022.3216084
12 W. Ma, K. Wang, J. Li, S. X. Yang, J. Li, L. Song, and Q. Li: Sensors **23** (2023) 599. https://doi.org/10.3390/s23020599
13 M. Silvia, S.-C. Huang, C. G. Dalla, R. d'Isa, V. Castoldi, E. Rossi, G. Comi, and L. Leocani: Front Neurosci. **16** (2022) 820155. https://doi.org/10.3389/fnins.2022.820155
14 L. Malasinghe, S. Katsigiannis, K. Dahal, and N. Ramzan: Expert Syst. Appl. **207** (2022) 117867. https://doi.org/10.1016/j.eswa.2022.117867
15 Y. Guo, Y. Yang, Y. Liu, Q. Li, F. Cao, M. Feng, H. Wu, W. Li, and Y. Kang: Electronics **10** (2021) 1278. https://doi.org/10.3390/electronics10111278
16 T. Bahar, P. Carolina, R. Rebecca, N. Rosemary, and R. Sundararaman, Z. Erin, M. Brice, V. René, S. H. Mostofsky: Biol. Psychiatry: Cognit. Neurosci. **6** (2021) 321. https://doi.org/10.1016/j.bpsc.2020.09.001
17 K. P. Ahlers, T. P. Gabrielsen, D. Lewis, A. M. Brady, and A. Litchford: School Psychology International **3** (2017) 586. https://doi.org/10.1177/0143034317719942
18 W.-T. Tsai, I. J. Lee, and C.-H. Chen, Univers. Access Inf. Soc. **20** (2021) 375. https://doi.org/10.1007/s10209-020-00724-9
19 J. Yu, H. Gao, J. Sun, D. Zhou, and Z. Ju: IEEE Trans. Cognit. Dev. Syst. **14** (2022) 1574. https://doi.org/10.1109/TCDS.2021.3124764
20 K. V. Sree and G. Jeyakumar: Proc. 2020 Computational Vision and Bio-Inspired Computing (ICCVBIC 2019) 355–363. https://doi.org/10.1007/978-3-030-37218-7_41
21 S. K. Yadav, A. Luthra, K. Tiwari, H. M. Pandey, and S. A. Akbar: Knowledge-Based Syst. **239** (2022) 107948. https://doi.org/10.1016/j.knosys.2021.107948
22 F. Negin, B. Ozyer, S. Agahian, S. Kacdioglu, and G. T. Ozyer: Neurocomputing **446** (2021) 145. https://doi.org/10.1016/j.neucom.2021.03.004
23 X. Liu, Y. Wu, M. Chen, T. Liang, F. Han, and X. Liu: Math. Biosci. Eng. **20** (2023) 8049. https://doi.org/10.3934/mbe.2023349
24 A. Zunino, P. Morerio, A. Cavallo, C. Ansuini, J. Podda, F. Battaglia, E. Veneselli, C. Becchio, and V. Murino: Proc. 2018 24th Int. Conf. Pattern Recognition (ICPR, 2018) 3421–3426. https://doi.org/10.1109/ICPR.2018.8545095
25 K. Sun, L. Li, L. Li, N. He, and J. Zhu: Proc. ICASSP 2020 - 2020 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP, 2020) 3387–3391. https://doi.org/10.1109/ICASSP40776.2020.9054641
26 C. Fan, B. Lu, J. Ren, F. Li, and M. Shi: J. Mech. Med. Biol. **22** (2022) 2240051. https://doi.org/10.1142/S0219519422400516
27 N. Martínez-Molina, S.-T. Siponkoski, and T. Särkämö: Ann. N. Y. Acad. Sci. **1515** (2022) 20. https://doi.org/10.1111/nyas.14800
28 S. M. Anzalone, J. Xavier, S. Boucenna, L. Billeci, A. Narzisi, F. Muratori, D. Cohen, and M. Chetouani: Pattern Recognit. Lett. **118** (2019) 42. https://doi.org/10.1016/j.patrec.2018.03.007
29 Y. Shi, W. Ren, W. Jiang, Q. Xu, X. Xu, and H. Liu: Proc. Intelligent Robotics and Applications (2022) 370–380. https://doi.org/10.1007/978-3-031-13844-7_36
30 M. R. Baihaqi, R. M. Negara, and R. Tulloh: Proc. 2022 5th Int. Seminar Research of Information Technology and Intelligent Systems (ISRITI, 2022) 93–98. https://doi.org/10.1109/ISRITI56927.2022.10052975
31 K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu: Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition (2020, CVPR) 180–189. https://doi.org/10.1109/CVPR42600.2020.00026
32 Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh: Proc. IEEE Trans. Pattern Analysis and Machine Intelligence (IEEE, 2021) 172–186. https://doi.org/10.1109/TPAMI.2019.2929257
33 J. Wang, S. Yan, Y. Xiong, and D. Lin: Proc. Computer Vision – ECCV 2020 (ECCV, 2020) 764–780. https://doi.org/10.1007/978-3-030-58601-0_45
34 P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, IEEE Trans. Pattern Anal. Mach. Intell. **41** (2019) 1963. https://doi.org/10.1109/TPAMI.2019.2896631
35 S. Kalikar, A. Sinha, S. Srivastava, and G. Aggarwal: Proc. Third Int. Conf. Communication, Computing and Electronics Systems (Singapore, 2022) 1015–1027. https://doi.org/10.1007/978-981-16-8862-1_66
36 G. O. Ribeiro, M. Grellert, and J. T. Carvalho: Proc. 2023 IEEE 36th Int. Symp. Computer-Based Medical Systems (CBMS, 2023) 225–230. https://doi.org/10.1109/CBMS58004.2023.00221
37 D. Neimark, O. Bar, M. Zohar, and D. Asselmann: Proc. 2021 IEEE/CVF Int. Conf. Computer Vision Workshops (ICCVW, 2021) 3156–3165. https://doi.org/10.1109/ICCVW54120.2021.00355
38 K. He, X. Zhang, S. Ren, and J. Sun: Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR, 2016) 770–778. https://doi.org/10.1109/CVPR.2016.90

39 Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo: Proc. 2021 IEEE/CVF Int. Conf. Computer Vision (ICCV, 2021) 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986
40 S. Ji, W. Xu, M. Yang, and K. Yu: IEEE Trans. Pattern Anal. Mach. Intell. **35** (2013) 221. https://doi.org/10.1109/TPAMI.2012.59
41 M. C. Leong, D. K. Prasad, Y. T. Lee, and F. Lin: Appl. Sci. **10** (2020) 557. https://doi.org/10.3390/app10020557

About the Authors

Hongwei Gao received his Ph.D. degree in the field of pattern recognition and intelligent system from Shenyang Institute of Automation (SIA), Chinese Academy of Sciences (CAS) in 2007. Since September 2015, he has been a professor of the School of Automation and Electrical Engineering, Shenyang Ligong University. Currently, he is the leader of academic direction for optical and electrical measuring technology and system. His research interests include digital image processing and analysis, stereo vision, and intelligent computation. He has published more than sixty technical papers in these areas as the first author or a coauthor. (ghw1978@sohu.com)

Xuna Wang received her B.S. degree in automation from Harbin University of Science and Technology, China, in 2020. She is working on her M.S. degree in intelligent systems at Shenyang Ligong University, China. Her research interests include human motion analysis, healthcare robotics, and human–computer interaction. (xuna_emmm666@163.com)

Zide Liu received his B.S. degree in automation from Northeastern University, China, in 2021. He is working on his M.S. degree in pattern recognition at Shenyang Ligong University, China. His research interests include pattern recognition, expression recognition, and medical robotics. (edam_lzd@163.com)

Yueqiu Jiang received her Ph.D. degree in computer application technology from Northeastern University, China, in 2004. Since 2010, she has been a full professor of Shenyang Ligong University. Currently, she is the leader of subject direction for signal and information processes. Her research interests include network management and image processing.
(yueqiujiang@sylu.edu.cn)