Human Factors in Design, Engineering, and Computing, Vol. 159, 2024, 878–888
https://doi.org/10.54941/ahfe1005654

DMGR: Divisible Multi-Complex Gesture Recognition Based on Word Segmentation Processing

Yuncheng Ge[1], Yewei Huang[2], Julei Ye[3], Huazixi Zeng[1], Hechong Su[1], and Zengyao Yang[4]

[1]Joint School of Design and Innovation, Xi'an Jiaotong University, Xi'an 710049, China
[2]Tongda College of Nanjing University of Posts and Telecommunications, Yangzhou 225127, China
[3]School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
[4]School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China

ABSTRACT

Gesture recognition is a critical component of human-computer interaction, with applications ranging from virtual reality to assistive technologies. However, recognizing complex, multi-complex gesture sequences remains challenging. This paper introduces DMGR, a novel algorithm for recognizing multi-complex dynamic gesture tasks. Utilizing a list model and word segmentation-like processing, DMGR parses complex gestures into constituent actions. The algorithm's core innovation lies in its contextual gesture recognition and probability density-based segmentation, significantly enhancing accuracy and efficiency. Empirical results demonstrate DMGR's effectiveness in advancing gesture recognition technology, particularly for intricate, multi-layered sequences. This research contributes to the evolving field of gesture-based interfaces, offering potential for more intuitive and sophisticated human-computer interaction systems.

Keywords: Dynamic gesture recognition, Human-computer interaction, Temporal logic relationship network

INTRODUCTION

Gesture recognition has become an integral component of human-computer interaction (HCI), with applications spanning virtual reality, assistive technologies, and robotic control systems. While traditional approaches have focused on recognizing isolated gestures, contemporary HCI applications often require the interpretation of complex gestural sequences. This paradigm shift necessitates systems capable of not only recognizing individual gestures but also accurately segmenting and interpreting sequences of complex gestures to infer user intent. Drawing parallels with natural language processing (NLP), where understanding complex sentences requires word segmentation and contextual comprehension, HCI faces similar challenges in multi-complex dynamic gesture interaction. The key lies in developing methods to accurately delineate individual gestures within a continuous sequence and establish contextual relationships between them.

This paper introduces DMGR (Divisible Multi-complex Gesture Recognition), a novel algorithm that addresses the existing and potential challenges. Inspired by word processing techniques in NLP, DMGR employs a list model to decompose complex gestural sequences into constituent operations and actions. The paper is structured as follows: First, it elucidates the algorithm framework for divisible multi-complex dynamic gesture task recognition and the underlying model based on word processing techniques. Subsequently, it provides a detailed exposition of the algorithm's implementation, encompassing feature extraction, gesture classification, segmentation, and optimization methodologies. Finally, the paper presents the experimental design and results, offering empirical validation of the proposed approach's efficacy.

The algorithm incorporates action elements to reduce gesture dimension and utilizes a probability density distribution-based segmentation technique for accurate partitioning. By enhancing recognition accuracy and reducing computational complexity, DMGR represents a significant advancement in gesture recognition, particularly for complex, multi-gesture sequences. This research paves the way for more sophisticated and intuitive human-computer interaction paradigms, contributing to the evolving landscape of gesture-based interfaces.

METHOD

Model of Divisible Multi-Complex Gesture Recognition (DMGR)

In the process of human-computer interaction (HCI), there are often many complex gesture operation tasks, and it is very difficult to recognize these complex gesture tasks directly. Therefore, this paper proposes a multi-complex dynamic gesture task recognition model similar to word processing, as shown in Figure 1. Compared with word segmentation in the field of NLP, it is considered that a multi-complex dynamic gesture task can be divided into a series of complex gesture operations, and each gesture can be divided into several continuous gesture actions, each action corresponds to a single gesture.

T is a given multi-complex dynamic gesture task, and V is the video input data recorded corresponding to the multi-complex dynamic gesture task T. V can represent a collection of a series of digital image frames, i.e

$$V = \{\mathrm{f_t} | \mathrm{t} = 1, 2, \cdots\}$$

Where $\mathrm{f_t}$ represents the image of frame t List is a list of consecutive elements in strict order. Each frame in V corresponds one by one to each element in the ListList (V) $= [\mathrm{f_1}, \mathrm{f_2}, \cdots, \mathrm{f_{t-1}}, \mathrm{f_t}, \mathrm{f_{t+1}}, \cdots], t = 1, 2, \cdots$

Based on the above principles, the following definitions are given:

Definition 1: A is a gesture composed of consecutive frames $\mathrm{f_t}$, P is the number of consecutive frames $\mathrm{f_t}$ in A, then there is

$$\mathrm{List}\,(A) = [f_t, f_{t+1}, \cdots, f_{t+p-1}], t = 1, 2, \cdots, P = 1, 2, \cdots$$

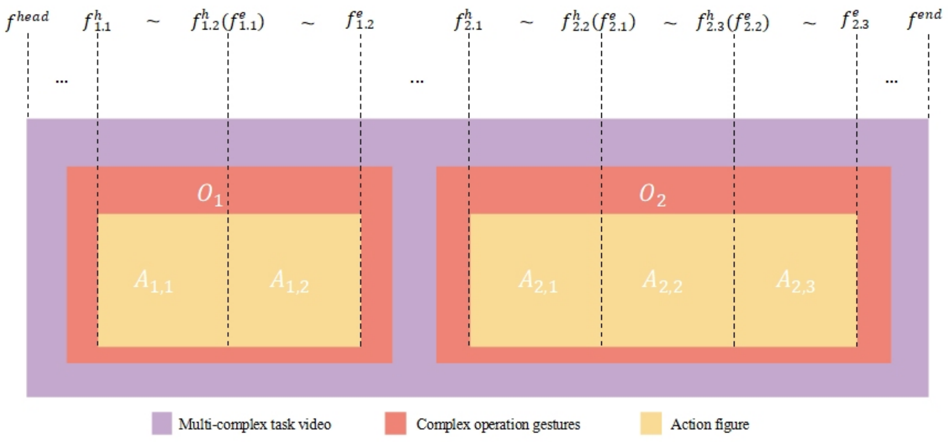

Figure 1: Model of multiple-complex gesture task recognition algorithm.

Definition 2: O is a complex gesture operation composed of multiple continuous gesture actions A, M is the number of continuous gesture actions A in O, then there is

$$\text{List}\,(O) = [A_1, A_2, \cdots, A_M], M = 1, 2, \cdots$$

Definition 3: T is a multi-complex dynamic gesture task composed of multiple continuous complex gesture operations O, N is the number of complex gesture operations O in T, then there is

$$\text{List}\,(T) = [\mathrm{O}_1, \mathrm{O}_2, \cdots, \mathrm{O}_N], N = 1, 2, \cdots$$

Definition 4: S is a cutting function that intercepts elements in the List between a given starting point and an ending point. $\mathrm{f}^{\mathrm{h}}_{\mathrm{i,j}}$ and $\mathrm{f}^{\mathrm{e}}_{\mathrm{i,j}}$ are multiple complex dynamic gestures task T, where the i complex gesture operates the first frame and the last of the j gesture action $A_{\mathrm{i.j}}$ in O_i, and there is

$$\mathrm{A}_{\mathrm{i.j}} = \mathrm{S}\left(\text{List}\,(\mathrm{V}), \mathrm{f}^{\mathrm{h}}_{\mathrm{i,j}}, \mathrm{f}^{\mathrm{e}}_{\mathrm{i,j}}\right) = [\mathrm{f}^{\mathrm{h}}_{\mathrm{i,j}}, \cdots, \mathrm{f}^{\mathrm{e}}_{\mathrm{i,j}}], \mathrm{i} = 1, 2, \cdots, \mathrm{N}, \mathrm{j} = 1, 2, \cdots, \mathrm{M}$$

Extraction of Multi-Complex Gesture Feature

For any frame image, the static feature matrix $\mathrm{F}^{\mathrm{Static}}_{\mathrm{t}}$ from f_1 to f_{t} can be defined as

$$F^{\text{Static}}_t = \begin{bmatrix} f^{\text{Static}}_1 \\ \underline{f^{\text{Static}}_2} \\ \vdots \\ f^{\text{Static}}_t \end{bmatrix}$$

$$f^{\text{Static}}_t = (Z^{\text{nm}}_t, V^{\text{HOG}}_t, f^{\text{Type}}_t)$$

Where f^{Static}_t is the static feature vector of f_t, Z^{nm}_t, V^{HOG}_t and f^{Type}_t are Zernike moments (Khotanzad and Hong, 1990) and directional gradient squares respectively Graph and gesture type characteristics.

This algorithm mainly uses Zernike moment (Khotanzad and Hong, 1990) and Histogram of Oriented Gradient (HOG) (Dalal and Triggs, 2005) to extract static representation features.

(1) Zernike moments

$\mathrm{p}_t^{'}(\mathrm{x_t}, \mathrm{y_t})$ is defined as a two-dimensional matrix of the normalized image f_t, and its Zernike moment of order n m is defined as:

$$\begin{aligned}
\mathrm{Z_t^{nm}} &= \frac{\mathrm{n}+1}{\pi}\sum_{\mathrm{x_t}}\sum_{\mathrm{y_t}}\mathrm{p}_t'(\mathrm{x_t},\mathrm{y_t})\mathrm{V_t^{nm*}}(\mathrm{x_t},\mathrm{y_t}),\ \mathrm{V_t^{nm*}}(\mathrm{x_t},\mathrm{y_t}) = \mathrm{R_t^{nm}}(\rho_\mathrm{t})\mathrm{e}^{-\mathrm{jm}\theta_\mathrm{t}} \\
\mathrm{R_t^{nm}}(\rho_\mathrm{t}) &= \sum_{\mathrm{s}=0}^{(\mathrm{n}-|\mathrm{m}|)/2}\frac{(-1)^\mathrm{s}\,(\mathrm{n}-\mathrm{s})\,!\,\rho^{\mathrm{n}-2\mathrm{s}}}{\mathrm{s}!\left(\frac{\mathrm{n}+|\mathrm{m}|}{2}-\mathrm{s}\right)!\left(\frac{\mathrm{n}-|\mathrm{m}|}{2}-\mathrm{s}\right)!} \\
\theta_\mathrm{t} &= \arctan(\mathrm{y_t}/\mathrm{x_t}), \rho_\mathrm{t} = \sqrt{\mathrm{x_t^2} + \mathrm{y_t^2}}
\end{aligned}$$

Where $*$ represents conjugations in complex numbers, $\mathrm{V_t^{nm}}$ is Zernike polynomial, $R_t^{\mathrm{nm}}(\rho_t)$ is radial polynomial, n and m are non-negative integers, and $m \leq n$, $n - m$ is odd, ρ_t is extreme, and θ_t is polar Angle.

(2) Histogram of Oriented Gradient (HOG)

The gradient of each pixel in the horizontal and vertical directions, as well as the gradient size and direction of each pixel position in the normalized input gesture image are calculated respectively:

$$\begin{aligned}
\mathrm{G_{tx}}(\mathrm{x_t},\mathrm{y_t}) &= \mathrm{p}_t'(\mathrm{x_t}+1,\mathrm{y_t}) - \mathrm{p}_t'(\mathrm{x_t}-1,\mathrm{y_t}) \\
\mathrm{G_{ty}}(\mathrm{x_t},\mathrm{y_t}) &= \mathrm{p}_t'(\mathrm{x_t},\mathrm{y_t}+1) - \mathrm{p}_t'(\mathrm{x_t},\mathrm{y_t}-1) \\
\nabla\mathrm{G_t}(\mathrm{x_t},\mathrm{y_t}) &= \sqrt{\mathrm{G_{tx}}(\mathrm{x_t},\mathrm{y_t})^2 + \mathrm{G_{ty}}(\mathrm{x_t},\mathrm{y_t})^2} \\
\theta_{\mathrm{tG}}(\mathrm{x_t},\mathrm{y_t}) &= \arctan\left(\mathrm{G_{ty}}(\mathrm{x_t},\mathrm{y_t})/\mathrm{G_{tx}}(\mathrm{x_t},\mathrm{y_t})\right)
\end{aligned}$$

Where $\mathrm{G_{tx}}(\mathrm{x_t},\mathrm{y_t})$, $\mathrm{G_{ty}}(\mathrm{x_t},\mathrm{y_t})$, $\nabla\mathrm{G}_t\left(x_t, y_t\right)$ and $\theta_{\mathrm{tG}}\left(x_t, y_t\right)$ are the gradient of pixel $\left(x_t, y_t\right)$ in the horizontal and vertical directions, and the gradient amplitude and gradient direction respectively.

(3) Dynamic feature extraction

For a frame of image, define $p(x_t^c,\ y_t^c)$ as the centroid point of the hand in image f_t. According to the continuous frames in the input video data V, two vectors can be constructed through the centroid point $p(x_t^c,\ y_t^c)$, i.e

$$X_t = (x_1^C, x_2^C, \cdots, x_t^C), Y_t = (y_1^C, y_2^C, \cdots, y_t^C)$$

Considering that the range and speed of gesture movement vary from person to person, the vectors X_t and Y_t are averaged to make them translation invariant and more stable. We then get two new vectors $X_t^{'}$ and $Y_t^{'}$:

$$X_t' = (x_1^C - \overline{X_t}, x_2^C - \overline{X_t}, \cdots, x_t^C - \overline{X_t}),\ Y_t' = (y_1^C - \overline{Y_t}, y_2^C - \overline{Y_t}, \cdots, y_t^C - \overline{Y_t})$$
$$\overline{X_t} = \frac{1}{t}\sum\nolimits_{t=1}^{t} x_i^C, \overline{Y_t} = \frac{1}{t}\sum\nolimits_{t=1}^{t} y_i^C$$

Where $\overline{X_t}$ and $\overline{Y_t}$ are the mean value of X_t and Y_t respectively. And the feature of dynamic gesture F_t^{Dyn} could be represented as:

$$F_t^{\mathrm{Dyn}} = \begin{bmatrix} X_t' \\ Y_t' \end{bmatrix}$$

Recognition and Classification of Divisible Multi-Complex Gesture

In order to improve the efficiency of gesture recognition, the concept of action element is introduced in this paper. L is defined as the length of the action element. The selection of L value should be neither too large nor too small. A large L will increase the computational complexity and affect the computational efficiency, while a small L will affect the recognition accuracy. In the study of this paper, the selection of the length L of the action element is verified through the experiment, and $L = 5$ is set.

The action element is defined as follows: for any frame image, f_t and its front row $L - 1$ frame constitute an action element U_t, namely

$$U_t = S\left(\mathrm{List}(V), f_{t-L+1}, f_t\right) = [f_{t-L+1}, f_{t-L+2}, \cdots, f_t]$$

Through the application of action elements, in the process of gesture recognition, the feature dimension can be reduced to L dimension. If the parameter L is reasonably selected, the computational complexity and time can be reduced without sacrificing the accuracy of the recognition algorithm. Moreover, the computational complexity of the recognition algorithm after the application of action elements is not greatly affected by the length of input video data, and even the input data of long and complex tasks can be recognized and classified quickly and effectively.

EXPERIMENTS

Divisible multi-complex gesture recognition based on word segmentation processing are presented in two public databases: Sheffield Kinect Gesture (SKIG) (Liu and Shao, 2013a) and Sebastien Marcel Dynamic Hand Posture (SMDHP) (Marcel et al., 2000) and a self-built database (Multi-complex Gesture). The algorithm proposed in this paper is compared on three data sets above, and the evaluation index is defined as segmentation accuracy (SA).

Experimental Setup

In each database, 20% of the data of various gestures were selected to train the template gesture feature sequence, and the remaining 80% data was used to test the accuracy of the algorithm. In the gesture recognition algorithm, the selection of the length L of the action element is verified through the experiment. According to the results, considering the trade-off of recognition accuracy and computational complexity, the length of action element $L = 30$ frames is set. In the gesture segmentation and optimization algorithm, the error threshold of $\mathrm{T}_1 = 0.3$ s and $T_2 = 2\ s$ is set through the statistics of the gesture duration in the commonly used gesture database.

The Evaluation Index

In this paper, a Segmentation Accuracy (SA) is defined as an index to evaluate the performance of the proposed segmentation algorithm. Segmentation accuracy refers to the percentage of overlap between experimental segmentation results and target segmentation results in target segmentation results, which is defined as follows:

$$SA = \frac{|\mathrm{F}_{\mathrm{result}} \cap \mathrm{F}_{\mathrm{groundtruth}}|}{\mathrm{F}_{\mathrm{groundtruth}}}$$

Where F_{result} is the experimental segmentation result and $F_{\mathrm{groundtruth}}$ is the target segmentation result. As can be seen from the formula, the higher the coincidence degree between the segmentation result obtained by the algorithm and the actual segmentation result, the greater the value of SA. The larger the SA value is, the higher the segmentation accuracy of the algorithm is. Otherwise, the lower the segmentation accuracy is.

Performance Comparison

(1) Sheffield Kinect Gesture (SKIG)

The experimental results on SKIG database are shown in the table. In order to verify the effectiveness of the algorithm, this paper compares it with RGGP + RGB-D (Liu and Shao, 2013a), 4DCOV (Cirujeda and Binefa, 2014), Depth Context (Liu, 2016), HOG + LBP (Azad et al., 2019) and DLEH2 (DLE+ HOG2) (Zheng et al., 2017).

Table 1. Recognition accuracy of SKIG database.

Method	Accuracy of Recognition (%)
4DCOV	93.2±0.6
Depth Context	95.2±1.9
HOG+LBP	97.3±1.6
DLEH2	98.2±1.1
RGGP+RGB-D	88.4±1.2
DMGR	99.2±0.6

The DMGR algorithm demonstrates superior performance in gesture recognition compared to HOG+LBP and DLEH2 methods. While HOG+LBP employs multi-level time sampling to improve intra-class similarity and inter-class differences, it lacks focus on spatio-temporal consistency, which is a key feature of DMGR. This consistency proves crucial in gesture recognition, as evidenced by DMGR's superior experimental results. Although DLEH2 integrates depth sequence gesture shape with temporal and spatial features, its reliance solely on HOG for feature extraction limits its ability to capture overall texture and edge information. In contrast, DMGR's comprehensive approach to feature learning and emphasis on spatio-temporal relationships enables it to more effectively recognize and interpret complex gesture sequences, resulting in higher accuracy and robustness in gesture recognition tasks.

(2) Sebastien Marcel Dynamic Hand Posture (SMDHP)

Discriminant Canonical Correlation Analysis (DCCA) (Kim, Kittler and Cipolla, 2007), Tensor Canonical Correlation Analysis (TCCA) (Kim, Wong and Cipolla, 2007), Product Manifold (PM) (Lui, Beveridge and Kirby, 2010), Genetic Programming (GP) (Liu and Shao, 2013b), Tangent Bundle (TB) (Lui, 2012) and 3D spatio-temporal covariance descriptors (Cov3D) (Sanin et al., 2013) and other algorithms are compared, and the comparison results are shown in the table.

Table 2. Recognition accuracy of SMDHP database.

Method	Accuracy of Recognition (%)
DCCA	65.5±5.5
TCCA	82.1±2.7
PM	88.4±3.4
GP	85.0±1.1
TB	91.7±0.9
Cov3D	93.3±1.2
DMGR	96.8±0.4

The proposed DMGR algorithm demonstrates superior recognition performance, surpassing DTW by up to 11.8%. Cov3D ranks second, effectively maintaining spatio-temporal consistency but lacking comprehensive gesture description. These results underscore the importance of balancing static features and dynamic spatio-temporal consistency in complex gesture recognition, highlighting DMGR's efficacy in capturing both aspects for improved accuracy.

(3) Multi-complex Gesture

On the self-built Multi-complex Gesture database, 50% cross validation is used, i.e. 80% of the data is used for training the model and the remaining 20% is used for validation. Table 3 shows the recognition results of the proposed algorithm on the Multi-complex Gesture database, and the comparison results between the proposed algorithm and SSBoW (Laptev and Lindeberg, 2005), DSBoW (Laptev et al., 2008), DTBoW (Hao et al., 2013; Wang et al., 2013), DFW (Kulkarni et al., 2015) and other algorithms.

Table 3. Recognition accuracy of multi-complex gesture database.

Motion	Point	Move	Scratch	Untie	Zoom	Rotate	None
Recognition	94.3	96.7	94.0	90.1	91.2	90.5	98.1
accuracy (%)	92.5	96.2	93.4	91.5	92.1	91.1	97.1
	91.6	95.3	92.1	90.4	91.5	91.3	95.4
	93.1	97.6	91.4	91.2	90.2	90.2	96.9
	92.8	95.3	92.1	91.1	92.3	91.2	97.9
Average	92.9±1.0	96.3±1.0	92.7±1.0	91.1±0.4	91.5±0.7	91.1±0.2	96.8±0.9

Table 4. Comparison of proposed algorithm with others in multi-complex gesture database.

Motion	Point	Move	Scratch	Untie	Zoom	Rotate	None
SSBoW	65.2±13.5	81.7±11.4	83.8±12.1	88.2±12.5	59.8±18.8	76.5±15.0	74.2±20.7
DSBoW	78.8±10.3	83.8±10.4	87.5±7.2	86.1±4.3	72.4±10.9	81.7±4.0	84.8±19.1
DTBoW	71.9±11.5	91.4±8.3	90.5±3.6	86.7±6.1	90.2±4.7	87.6±3.5	91.1±9.2
DFW	85.8±6.2	94.0±2.3	90.9±7.8	88.3±2.6	88.9±3.4	85.1±2.3	93.2±5.8
DMGR	92.9±1.0	96.3±1.0	92.7±1.0	91.1±0.4	91.5±0.7	91.1±0.2	96.8±0.9

It can be seen that the recognition accuracy of each type of gesture in the proposed algorithm is above 90%, and the overall average recognition rate is 92.9%. Compared with other algorithms, the proposed algorithm also achieves the highest recognition accuracy, especially for the two gestures of pointing and zooming, which is much higher than other algorithms. Figure 2 shows the frame-to-frame recognition results of each comparison algorithm.

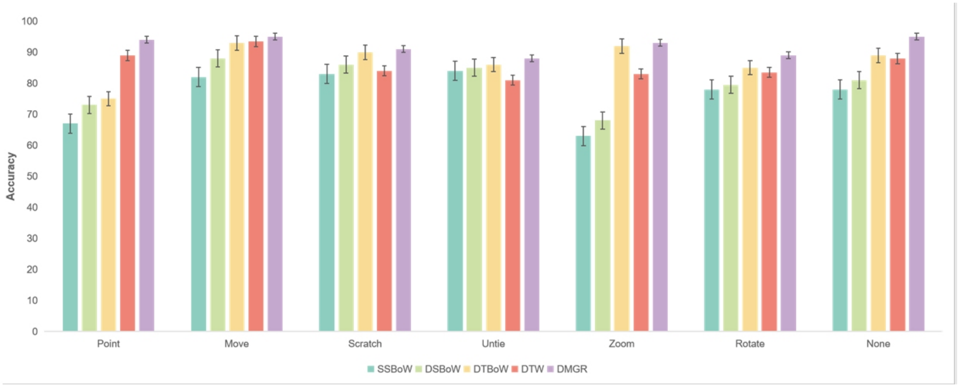

Figure 2: Accuracy of frames recognition on multi-complex gesture database.

In order to verify the effectiveness of the segmentation algorithms, the segmentation performance of different algorithms is evaluated and compared, as shown in Figure 2. The DMGR algorithm demonstrates superior segmentation performance compared to SSBoW, DSBoW, DTBoW, and DFW methods. Its effectiveness is attributed to the accurate identification of meaningful gestures and iterative optimization of segmentation boundaries. This precise segmentation contributes to the algorithm's higher recognition accuracy. In Figure 3, experiments with varying image resolutions reveal that DMGR maintains over 90% accuracy for resolutions above 160×120. Performance declines for lower resolutions, with 64% accuracy at 40×30. However, such low-resolution inputs are uncommon in practical applications, suggesting the algorithm's reliability under most real-world conditions. These findings validate DMGR's effectiveness in gesture segmentation and recognition across diverse camera conditions, underscoring its potential for widespread application in gesture-based interfaces.

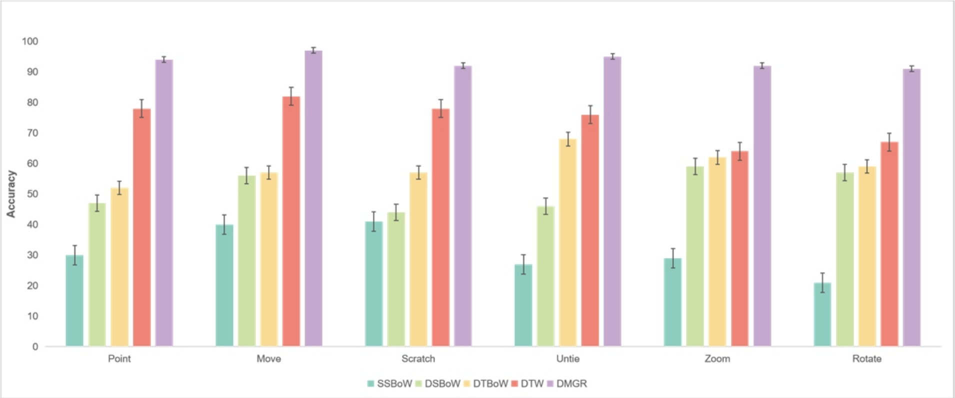

Figure 3: Accuracy of action boundary segmentation on multi-complex gesture database.

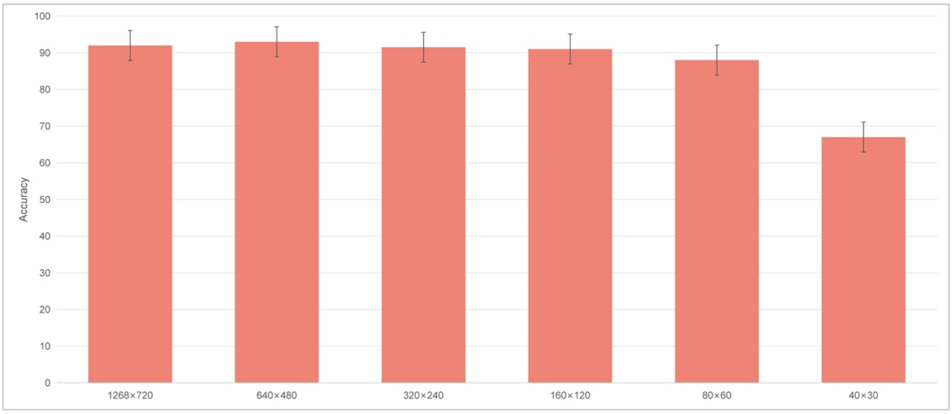

Figure 4: Effect of image resolution on accuracy.

CONCLUSION

This study presents a novel algorithm for recognizing multi-complex dynamic gesture tasks. A key innovation of this research is the development of a gesture segmentation and optimization algorithm founded on probability density distribution. This component significantly enhances the accuracy of gesture segmentation within multi-complex gesture tasks, addressing a critical challenge in gesture recognition systems. Empirical evaluation of the proposed algorithm demonstrates its efficacy, revealing substantial improvements in both recognition accuracy and computational efficiency when processing multi-complex gesture tasks.

In conclusion, this research presents a robust and efficient solution to the challenging problem of multi-complex dynamic gesture task recognition, paving the way for more sophisticated and natural gesture-based interaction systems in various technological domains.

DISCLOSURE STATEMENT

The authors report there are no competing interests to declare.

ACKNOWLEDGMENT

This research was funded by Provincial Training Program of Innovation and Entrepreneurship for Undergraduates.

REFERENCES

Azad, R., Asadi-Aghbolaghi, M., Kasaei, S. and Escalera, S. (2019) 'Dynamic 3D Hand Gesture Recognition by Learning Weighted Depth Motion Maps', *IEEE Transactions on Circuits and Systems for Video Technology*, 29(6), pp. 1729–1740. Available at: https://doi.org/10.1109/TCSVT.2018.2855416.

Cirujeda, P. and Binefa, X. (2014) '4DCov: A Nested Covariance Descriptor of Spatio-Temporal Features for Gesture Recognition in Depth Sequences', in *2014 2nd International Conference on 3D Vision. 2014 2nd International Conference on 3D Vision (3DV)*, Tokyo: IEEE, pp. 657–664. Available at: https://doi.org/10.1109/3DV.2014.10.

Dalal, N. and Triggs, B. (2005) 'Histograms of Oriented Gradients for Human Detection', in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA: IEEE, pp. 886–893. Available at: https://doi.org/10.1109/CVPR.2005.177.

Hao, Z., Zhang, Q., Ezquierdo, E. and Sang, N. (2013) 'Human action recognition by fast dense trajectories', in *Proceedings of the 21st ACM international conference on Multimedia. MM '13: ACM Multimedia Conference*, Barcelona Spain: ACM, pp. 377–380. Available at: https://doi.org/10.1145/2502081.2508123.

Khotanzad, A. and Hong, Y. H. (1990) 'Invariant image recognition by Zernike moments', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5), pp. 489–497. Available at: https://doi.org/10.1109/34.55109.

Kim, T.-K., Kittler, J. and Cipolla, R. (2007) 'Discriminative Learning and Recognition of Image Set Classes Using Canonical Correlations', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6).

Kim, T.-K., Wong, S.-F. and Cipolla, R. (2007) 'Tensor Canonical Correlation Analysis for Action Classification', in *2007 IEEE Conference on Computer Vision and Pattern Recognition. 2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA: IEEE, pp. 1–8. Available at: https://doi.org/10.1109/CVPR.2007.383137.

Kingma, D. P. and Ba, J. (2014) 'Adam: A Method for Stochastic Optimization', *CoRR* [Preprint]. Available at: https://www.semanticscholar.org/paper/a6cb366736791bcccc5c8639de5a8f9636bf87e8 (Accessed: 18 July 2024).

Kulkarni, K., Evangelidis, G., Cech, J. and Horaud, R. (2015) 'Continuous Action Recognition Based on Sequence Alignment', *International Journal of Computer Vision*, 112(1), pp. 90–114. Available at: https://doi.org/10.1007/s11263-014-0758-9.

Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B. (2008) 'Learning realistic human actions from movies', in *2008 IEEE Conference on Computer Vision and Pattern Recognition. 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, AK, USA: IEEE, pp. 1–8. Available at: https://doi.org/10.1109/CVPR.2008.4587756.

Liu, L. and Shao, L. (2013a) 'Learning Discriminative Representations from RGB-D Video Data', in *In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. Twenty-Third international joint conference on Artificial Intelligence*, AAAI Press.

Liu, L. and Shao, L. (2013b) 'Synthesis of spatio-temporal descriptors for dynamic hand gesture recognition using genetic programming'.

Liu, M. (2016) 'Depth Context_ a new descriptor for human activity recognition by using sole depth sequences'.

Lui, Y. M. (2012) 'Tangent Bundles on Special Manifolds for Action Recognition', *IEEE Transactions on Circuits and Systems for Video Technology*, 22(6), pp. 930–942. Available at: https://doi.org/10.1109/TCSVT.2011.2181452.

Lui, Y. M., Beveridge, J. R. and Kirby, M. (2010) 'Action classification on product manifolds', in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA: IEEE, pp. 833–839. Available at: https://doi.org/10.1109/CVPR.2010.5540131.

Marcel, S., Bernier, O., Viallet, J.-E. and Collobert, D. (2000) 'Hand gesture recognition using input-output hidden Markov models', in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580). Fourth International Conference on Automatic Face and Gesture Recognition*, Grenoble, France: IEEE Comput. Soc, pp. 456–461. Available at: https://doi.org/10.1109/AFGR.2000.840674.

Sanin, A., Sanderson, C., Harandi, M. T. and Lovell, B. C. (2013) 'Spatio-temporal covariance descriptors for action and gesture recognition', in *2013 IEEE Workshop on Applications of Computer Vision (WACV). 2013 IEEE Workshop on Applications of Computer Vision (WACV)*, Clearwater Beach, FL, USA: IEEE, pp. 103–110. Available at: https://doi.org/10.1109/WACV.2013.6475006.

Wang, H., Kläser, A., Schmid, C. and Liu, C.-L. (2013) 'Dense Trajectories and Motion Boundary Descriptors for Action Recognition', *International Journal of Computer Vision*, 103(1), pp. 60–79. Available at: https://doi.org/10.1007/s11263-012-0594-8.

Zheng, J., Feng, Z., Xu, C., Hu, J. and Ge, W. (2017) 'Fusing shape and spatio-temporal features for depth-based dynamic hand gesture recognition', *Multimedia Tools and Applications*, 76(20), pp. 20525–20544. Available at: https://doi.org/10.1007/s11042-016-3988-8.