A frame-level video annotation tool for dynamic gestures and apply in Vietnamese sign language

Nguyen Trung Hieu[1], Le Dinh Anh[1], Duong Van An[1], Pham Thi Thanh Thuy[2], Pham Ngoc Khiem[3], Tran Quang Truong[1], Trinh Thi Hoang[1], Doan Thi Huong Giang[1*]

[1]Faculty of Control and Automation, Electric Power University, 235 Hoang Quoc Viet, Nghia Do, Hanoi, Vietnam;
[2]Faculty of Cybersecurity and High Tech Crime Prevention, Academy of People Security, 125 Tran Phu, Ha Dong, Hanoi, Vietnam;
[3]School of Electrical Engineering, Hanoi University of Science and Technology, 1 Dai Co Viet, Bach Mai, Hanoi, Vietnam.
[*]Corresponding author: giangdth@epu.edu.vn
Received 28 Mar. 2026; Revised 1 Jun. 2026; Accepted 15 Jun. 2026; Published 25 Jun. 2026.
DOI: https://doi.org/10.54939/1859-1043.j.mst.112.2026.167-175

ABSTRACT

Human Action Recognition (HAR) in video is essential for human–computer interaction, particularly in sign language and smart device control. However, model performance depends heavily on accurate temporal annotation of dynamic gestures. This study proposes a frame-level video annotation tool for Vietnamese sign language, enabling precise temporal segmentation and structured JSON export for deep learning applications. A dataset of 15 dynamic gesture classes is also constructed using a multi-view acquisition setup. Annotation quality is evaluated using Temporal IoU, Cohen's Kappa, and annotation time. Results show high inter-annotator agreement at 0.9470 and 0.9888, respectively, demonstrating the effectiveness of the proposed tool for reliable and efficient gesture annotation.

Keywords: Dynamic action recognition; Deep learning; Video annotation; Temporal IoU; Cohen's Kappa; Temporal segmentation; Annotation evaluation.

1. INTRODUCTION

Computer vision and deep learning have significantly advanced human–machine interaction, particularly in human action recognition and sign language recognition [1-4]. Large-scale datasets such as Kinetics [2], ActivityNet [4], and Pascal VOC [5] have accelerated this progress. However, performance remains highly dependent on annotation quality, especially the accurate temporal boundaries of dynamic gestures [4].

Existing datasets are often based on pre-segmented clips [6, 7], which do not fully represent continuous real-world actions. Therefore, accurate frame-level annotation of raw videos is essential for building reliable continuous gesture datasets. Although ActivityNet [4] and HACS [9] support temporal action annotation, they focus on generic actions rather than sign language, where gesture boundaries may be ambiguous and sign order may differ from spoken-language structure.

Although VIA [8], CVAT [10], Label Studio [12], and ELAN [13] support general annotation tasks, they are not specifically designed for sentence-level sign language videos. VIA, CVAT, and Label Studio provide limited frame-level temporal segmentation and lack sentence–gloss order mapping, while ELAN's tier-based structure can be cumbersome and often requires additional processing for deep-learning formats. Therefore, these tools are less suitable for Vietnamese sign language, where sign order may differ from spoken-language word order. Annotation quality is assessed using tIoU [4], Cohen's Kappa [11], and annotation time to measure temporal overlap, label agreement, and efficiency. Since Kappa may be affected by class imbalance [14], these metrics provide complementary evaluation.

We introduce a Vietnamese sign language dataset with 15 dynamic gesture classes collected

using a multi-view setup. Together with the proposed annotation tool, it supports precise temporal segmentation and structured sentence-level annotation. The main contributions are: (1) A Vietnamese sign language gesture dataset for human–machine interaction; (2) A frame-level annotation tool with structured data export; (3) An evaluation framework using tIoU, Cohen's Kappa, and annotation time.

Section 2 presents the method, Section 3 the evaluation metrics, Section 4 the dataset, and Section 5 the experimental results and discussion.

2. PROPOSED METHOD

To support gesture recognition training, we develop a dedicated annotation tool for accurate frame-level segmentation and gesture labeling in raw videos.

2.1. Software pipeline

The proposed annotation process is organized as a multi-stage pipeline that transforms raw video data into structured annotations for training artificial intelligence models, as illustrated in Figure 1.

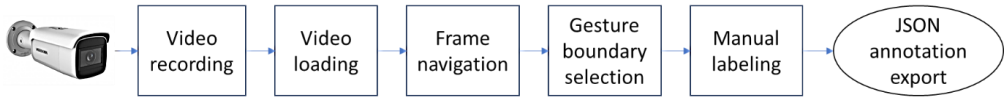


Figure 1. *Pipeline of the proposed video annotation process.*

The pipeline comprises multi-view video acquisition, frame-level visualization and navigation, gesture boundary identification, gesture labeling, and JSON export. It supports accurate temporal annotation and produces structured data for deep learning and gesture recognition. Figure 2 presents the corresponding pseudo-code.

```
Algorithm: Frame-Level Annotation Procedure for Vietnamese Sign Language Videos
Require: Raw video V, label set C, full sentence gloss F, frame rate r
Ensure: Structured annotation file A in JSON format
1: Load video V and extract metadata including frame rate r and total number of
frames N
2: Initialize an empty annotation list A
3: Display V with frame-level navigation controls
4: while there are unannotated gesture segments in V do
5: Navigate through V to identify the start frame fs of a gesture segment
6: Navigate through V to identify the end frame fe of the gesture segment
7: Select the corresponding gesture label c from C
8: Enter the full sentence gloss F in spoken-language order
9: if fs < fe and c is valid then
10: Compute the segment duration d = (fe - fs) / r
11: Append record (fs, fe, c, F, d) to A
12: else
13: Request correction of the invalid annotation
14: end if
15: end while
16: Review, edit, or delete annotation records if necessary
17: Export A as a JSON file
18: return A
```

Figure 2. *Pseudo-code of the proposed frame-level annotation procedure.*

2.2. Annotation software development

The annotation software is developed using Python. The main objective of this tool is to provide fine-grained control over individual video frames, which is difficult to achieve with conventional video playback software when annotating dynamic gestures. The user interface is designed with the following main components:

- Video display and control panel: The system allows users to upload and visualize videos. To accurately determine gesture timing, navigation controls are provided to move forward or backward by small units, including single frames or blocks of 10 frames. Additionally, a progress slider enables quick navigation to desired positions in the video.

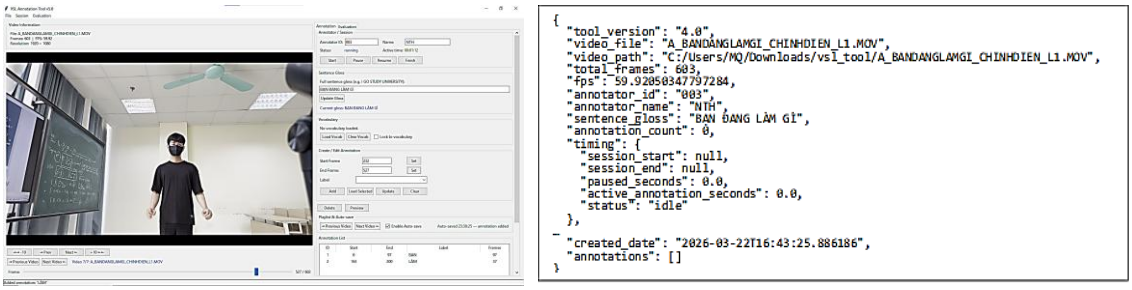
Input panel: This section allows users to input annotation information. A gesture instance c annotated by annotator i is defined in Equation (1):

$$G_c^i = \left(f_s^i, f_e^i, L^i\right) \quad (1)$$

Where f_s^i is starting frame labeled by the i[th] annotator, f_e^i is ending frame labeled by the i[th] annotator, and L^i: gesture label assigned by the i[th] annotator.

Annotation list management panel: All created annotations are displayed in a tabular list. Users can review, edit, or delete annotations before finalizing the dataset, ensuring annotation accuracy.

Figure 3 (a) illustrates the interface of the software developed for video annotation.



a) *b)*

Figure 3. *Illustration of the user interface of the video annotation tool (a) and record in the JSON file (b).*

2.3. Output data structure

After annotation, the data are exported in JSON format for seamless integration with machine learning frameworks. The file includes video information, metadata (frame count and frame rate), and a list of annotated gestures. Each annotation record contains the start frame, end frame, gesture label, and duration. An example of the JSON record is shown in Figure 3 (b).

Storing frame-level boundaries enables precise extraction of gesture segments while removing irrelevant segments such as idle states or noise. This results in cleaner datasets suitable for training AI models.

3. EVALUATION METRICS

In this study, three metrics are used to evaluate the annotation tool: Cohen's Kappa (K), Temporal IoU (Intersection over Union - t_{IoU}) and Annotation Time.

3.1. Temporal IoU

Temporal IoU is used to measure the degree of overlap between two temporal segments of a gesture annotated by two different annotators. Let the dataset contain C = {c | c = (1, ..., C)} gesture classes and a total of N = {n | n = (1, ..., N)} gesture instances. For a given gesture instance n belonging to class c = L, when it is annotated by annotator i, it is represented by the following parameters as defined in Equation (2):

$$G_c^i = [f_s^i, f_e^i, L^i) \quad (2)$$

The same gesture instance n, belonging to class c = L, is also annotated by annotator j and represented by the parameters defined in Equation (3):

$$G_c^j = [f_s^j, f_e^j, L^j) \quad (3)$$

If $L^i = L^j = L$ the Temporal IoU metric t_{IoU}^L for gestures belonging to class c=L annotated by two annotators, is computed based on the gesture annotated by annotator i[th] (G^i) and the gesture annotated by annotator j[th] (G^j) as defined in Equation (4):

$$t_{IoU}^n = \frac{t_{Intersection}(G^i, G^j)}{t_{Union}(G^i, G^j)} = \frac{max\left[0, min\left(f_e^i, f_e^j\right) - max\left(f_s^i, f_s^j\right)\right]}{max\left(f_e^i, f_e^j\right) - min\left(f_s^i, f_s^j\right)} \quad (4)$$

Where $t_{Intersection}(G^i, G^j)$ denotes the overlapping temporal duration between the two annotated gestures, while $t_{Union}(G^i, G^j)$ represents the total temporal span covered by both gestures. For a dataset containing N annotated gesture instances, the overall Temporal IoU is computed as defined in Equation (5):

$$t_{IoU} = \frac{1}{N}\sum_{n=1}^{N} t_{IoU}^{n} \tag{5}$$

The value of t_{IoU} ranges from 0 to 1. A lower t_{IoU} indicates a smaller overlap between annotated gestures, whereas a value approaching 1 suggests that the two annotators have identified nearly identical temporal boundaries for the gesture.

3.2. Cohen's Kappa metric

In addition to evaluating the temporal boundary overlap of gestures using Temporal IoU (t_{IoU}), this study also employs a second evaluation metric, Cohen's Kappa (K), to measure the agreement of gesture labels c = L between two annotators. Unlike simple agreement ratios, Cohen's Kappa accounts for the possibility that two annotators may assign the same label purely by chance. This metric is therefore used to assess the true level of agreement between annotators after removing the effect of random agreement. In some cases, annotators may coincidentally assign identical labels; thus, Cohen's Kappa provides a more accurate reflection of the actual agreement between annotators.

Assume that two annotators i and j label the same dataset as described in Section 3(a). The label assigned to the n-th gesture by the two annotators is denoted as L_n^i and L_n^j, respectively. The observed agreement ratio, or actual agreement (P_0) between the two annotators is computed as defined in Equation (6):

$$P_0 = \frac{1}{N}\sum_{n=1}^{N} 1\left(L_n^i = L_n^j\right) \tag{6}$$

While 1(·) denotes the indicator function, which takes the value 1 if the two annotators assign the same label and 0 otherwise. Let p_c^i be the probability that annotator i assigns label c=L_n^i, and p_c^j be the probability that annotator j assigns label c=L_n^j. The expected agreement due to chance, P_e is computed as defined in Equation (7):

$$P_e = \sum_{c=1}^{C} p_c^i \,.\, p_c^j \tag{7}$$

Cohen's Kappa (K) for the entire dataset consisting of N gesture instances is then computed as defined in Equation (8) as follows:

$$K = \frac{P_0 - P_e}{1 - P_e} \tag{8}$$

A lower value of K indicates weaker agreement between annotators, whereas a value approaching 1 signifies a high level of agreement.

3.3. Annotation time metric

Annotation time is also used to evaluate tool efficiency, based on the actual time required to annotate a video or video set. Lower annotation time indicates improved efficiency and is computed using Equation (9):

$$T_{Annotation} = \frac{1}{N}\sum_{n=1}^{N} T_n \tag{9}$$

While $T_{Anotation}$ denotes the average annotation time, T_n is the time required to annotate the n-th video belonging to class c, N is the total number of videos in the dataset, and C is the total

number of gesture classes used in the evaluation. A smaller value of $T_{Anotation}$ indicates that the annotation tool is more effective in improving annotator efficiency. While the previous two metrics evaluate agreement between annotators, this metric is assessed independently and is mainly used to compare annotators or annotation tools. A lower average annotation time suggests that the tool better supports reducing annotation effort. This metric is typically used to compare different annotation tools when performed by the same annotator, rather than comparing multiple annotators on the same dataset.

In the experiments, annotators independently annotate the same set of videos in order to collect data for computing the three evaluation metrics.

4. DATASET

4.1. Dataset definition

In this study, the data sample space is defined as a finite set of commands and phrases in Vietnamese sign language, designed for human–machine interaction tasks. The dataset is structured into three main functional groups: (i) status communication, (ii) smart device control, and (iii) navigation. In total, the dataset consists of 15 labels, corresponding to 15 standardized commands or phrases, which are categorized based on their functional and semantic characteristics, as described in Table 1:

Table 1. *Definition of hand gesture classes.*

No.	Label	Functional category	Word order	Semantic description
1	I am cooking	Status & Feedback	I – COOK – EAT – PROG	Indicating a personal activity
2	I am working	Status & Feedback	I – WORK – PROG	Indicating a personal activity
3	I agree	Status & Feedback	I – AGREE	Confirming agreement
4	I disagree	Status & Feedback	I – AGREE – NOT	Confirming rejection
5	I understand	Status & Feedback	I – UNDERSTAND	Confirming comprehension
6	I do not understand	Status & Feedback	I – UNDERSTAND – NOT	Indicating lack of comprehension
7	Turn on the light	Device Control	LIGHT – ON	Activating an electrical device
8	Turn off the light	Device Control	LIGHT – OFF	Deactivating an electrical device
9	Open the door	Device Control	DOOR – OPEN	Controlling access (entry/exit)
10	Close the door	Device Control	DOOR – CLOSE	Controlling access (entry/exit)
11	Turn right	Navigation	DIRECTION – RIGHT – TURN	Directional movement command
12	Turn left	Navigation	DIRECTION – LEFT – TURN	Directional movement command
13	Go straight	Navigation	DIRECTION – STRAIGHT – GO	Forward movement command
14	Step back	Navigation	DIRECTION – BACK – GO	Backward movement command
15	Stop	Navigation	GO – STOP	Emergency stop command

4.2. Data specification

All labels correspond to dynamic gestures represented as temporal action sequences. Each gesture is defined by three components: hand shape, position and orientation relative to the frontal camera, and motion trajectory. The gestures are designed based on common communication conventions in the Vietnamese deaf community, ensuring practical relevance.

4.3. Data collection environment setup

The data were collected using a multi-view recording system to ensure comprehensive spatial information of the gestures. The system combines an IPC-A32P-PRO surveillance camera and a smartphone camera (iPhone) to diversify the input data sources. The acquisition setup consists of three cameras arranged in a multi-view configuration as illustrated in Figure 4, including one frontal camera and two additional cameras positioned at left and right viewpoints with an angle of 30^0. This configuration helps reduce occlusion and enhances the capture of hand gesture information during data collection.

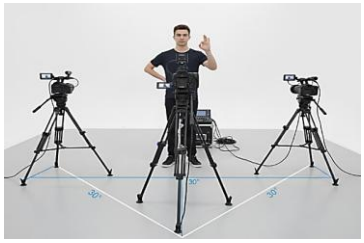


Figure 4. *Illustration of the data collection environment.*

The distance between the camera system and the subject is set within the range of 1.5 m to 2.0 m. All cameras are fixed throughout the data acquisition process to ensure consistency in viewpoint and perspective.

4.4. Data collection procedure

Each data sample is collected following a three-stage procedure:(1) Resting state: the subject remains in a natural initial posture; (2) Gesture execution: the subject performs the complete gesture corresponding to the assigned label at a natural speed; (3) Completion: the subject returns to the resting state to clearly define the temporal boundaries of the gesture. Detailed descriptions of how each gesture is performed and assigned EPUVSLVideo. This dataset was provided in Table 1.

To enhance generalization capability, the dataset is collected with controlled variations, including: (1) slight changes in body posture and orientation; (2) minor variations in position and distance relative to the camera; (3) diversity in clothing and physical appearance. These factors enable the model to learn robust gesture representations and reduce its dependency on specific data acquisition conditions.

5. EXPERIMENTAL RESULT

Experiments were implemented in Python on an Intel Core i5-11400H CPU, NVIDIA GTX 1650 GPU, and 8 GB RAM. Evaluation used EPUVSLVideo (45 videos, 15 classes) and an IPN Hand subset [15] (45 videos, 68 segments). Four annotators independently labeled both datasets, yielding six pairwise comparisons assessed by mean tIoU, Cohen's Kappa, and annotation time.

5.1. Evaluation of overall agreement and pairwise analysis

Table 4 shows high agreement on EPUVSLVideo, with an average tIoU of 0.9470 and Cohen's Kappa of 0.9888. A1–A2 achieved the highest tIoU (0.9797), while A3–A4 obtained the lowest (0.9260), which still indicates strong temporal consistency. Evaluation on the IPN Hand subset further confirms that the proposed workflow generalizes to other continuous hand gesture datasets under different categories and recording conditions.

Table 2. *Pairwise evaluation of annotation agreement on EPUVSLVideo and IPN Hand subset.*

Annotator Pair	tIoU		Cohen's Kappa	
	EPUVSLVideo	**IPN Hand subset[15]**	**EPUVSLVideo**	**IPN Hand subset[15]**
A1–A2	0.9797 ± 0.0209	0.9724 +/- 0.0268	0.9778 ± 0.1491	0.9778 +/- 0.1491
A1–A3	0.9483 ± 0.0401	0.9462 +/- 0.0417	0.9778 ± 0.1491	0.9778 +/- 0.1491
A1–A4	0.9492 ± 0.1484	0.9448 +/- 0.0526	0.9773 ± 0.1508	0.9556 +/- 0.2084
A2–A3	0.9408 ± 0.0468	0.9389 +/- 0.0499	1.0000 ± 0.0000	1.0000 +/- 0.0000
A2–A4	0.9378 ± 0.1479	0.9356 +/- 0.0574	1.0000 ± 0.0000	0.9778 +/- 0.1491
A3–A4	0.9260 ± 0.1466	0.9284 +/- 0.0712	1.0000 ± 0.0000	1.0000 +/- 0.0000
Average	**0.9470 ± 0.0182**	0.9444 +/- 0.0152	**0.9888 ± 0.0113**	0.9815 +/- 0.0167

Table 3 further confirms the consistency across annotators. All pairwise tIoU values exceed 0.92, indicating no substantial disagreement in temporal boundary selection. Similarly, several pairs achieve perfect Kappa values of 1.0000, reflecting near-complete agreement in label assignment.

Table 3. *Pairwise annotation agreement matrix on EPUVSLVideo (tIoU/Cohen's Kappa).*

	A1	**A2**	**A3**	**A4**
A1	–	0.980 / 0.978	0.948 / 0.978	0.949 / 0.977
A2	0.980 / 0.978	–	0.941 / 1.000	0.938 / 1.000
A3	0.948 / 0.978	0.941 / 1.000	–	0.926 / 1.000
A4	0.949 / 0.977	0.938 / 1.000	0.926 / 1.000	–

Overall, these results indicate that annotation variability mainly arises from temporal boundary ambiguity rather than label interpretation. The consistently high agreement further demonstrates that the proposed annotation guidelines and software effectively support reliable and consistent annotation performance.

5.2. Per-class evaluation of annotation agreement

The aggregated results in Table 4 show that most gesture classes achieve a high level of annotation agreement. Specifically, 11 out of 15 classes obtain Mean tIoU values above 0.95 with low standard deviation, while Cohen's Kappa reaches 1.0000 for nearly all classes, indicating highly consistent label assignment.

However, several classes exhibit lower agreement. Notably, DONG_CUA shows the lowest Mean $\boldsymbol{t_{IoU}}$ at 0.7840 ± 0.2056, reflecting significant variation in temporal boundary selection, while DI_THANG presents a lower Kappa value of 0.8333 ± 0.2887, indicating inconsistencies in label assignment for certain samples.

Overall, these results suggest that annotation variability mainly arises from temporal boundary ambiguity rather than label misunderstanding. At the same time, the consistently high agreement across most gesture classes demonstrates that the proposed annotation software effectively supports annotators in achieving reliable and consistent labeling performance.

Table 4. *Aggregated annotation agreement results by gesture class on EPUVSLVideo.*

Gesture label	Video number	t_IoU	Kappa
BAT_DEN	3	0.9506 ± 0.0336	1.0000 ± 0.0000
DI_THANG	3	0.9542 ± 0.0265	0.8333 ± 0.2887
DONG_CUA	3	0.7840 ± 0.2056	1.0000 ± 0.0000
DUNG_LAI	3	0.9401 ± 0.0564	1.0000 ± 0.0000
LUI_BUOC	3	0.9810 ± 0.0124	1.0000 ± 0.0000
MO_CUA	3	0.9585 ± 0.0376	1.0000 ± 0.0000
RE_PHAI	3	0.9648 ± 0.0179	1.0000 ± 0.0000
RE_TRAI	3	0.9812 ± 0.0047	1.0000 ± 0.0000
TAT_DEN	3	0.9799 ± 0.0050	1.0000 ± 0.0000
TOI_DANG_LAM_VIEC	3	0.9546 ± 0.0210	1.0000 ± 0.0000
TOI_DANG_NAU_AN	3	0.9595 ± 0.0196	1.0000 ± 0.0000
TOI_DONG_Y	3	0.9653 ± 0.0254	1.0000 ± 0.0000
TOI_HIEU	3	0.9361 ± 0.0324	1.0000 ± 0.0000
TOI_KHONG_DONG_Y	3	0.9387 ± 0.0186	1.0000 ± 0.0000
TOI_KHONG_HIEU	3	0.9561 ± 0.0242	1.0000 ± 0.0000

5.3. Annotation time evaluation

The results in Table 5 show noticeable differences in annotation time among annotators. Specifically, annotator A1 achieves the lowest average time at 26.47 s/video and 18.05 s/segment, while A3 requires the longest time at 57.67 s/video and 39.32 s/segment. The overall average annotation time is 38.95 s/video and 26.63 s/segment.

Table 5. *Annotation time statistics across annotators on EPUVSLVideo.*

Annotator	Videos	Time (s)	Time/vid (s)	Segments	Time/seg (s)
A1	45	1191.25	26.47	66	18.05
A2	45	1905.56	42.35	66	28.87
A3	45	2595.25	57.67	66	39.32
A4	45	1319.06	29.31	65	20.29
Avr ± STD		**1752.78 ± 533.02**	**38.95 ± 12.50**	**65.75 ± 0.50**	**26.63 ± 9.33**

These differences suggest variability in annotation speed, likely due to differences in experience, familiarity with the tool, or individual annotation strategies. However, when considered alongside the high agreement results reported earlier, it can be observed that annotation quality remains consistently high despite variations in annotation time. This indicates that the proposed annotation tool provides stable support for annotators, enabling reliable and consistent labeling performance regardless of individual annotation speed.

6. DISCUSSION AND CONCLUSIONS

The proposed tool achieved high inter-annotator agreement on EPUVSLVideo, with a mean tIoU of 0.9470 and Cohen's Kappa of 0.9888, while evaluation on the IPN Hand subset confirmed its generalizability. Disagreements mainly arose from ambiguous temporal boundaries rather than gesture labels. The tool supports frame-level segmentation and sentence gloss input, making it suitable for Vietnamese sign language videos with differing sign and spoken-language orders. Future work will expand the dataset and annotator pool and investigate AI-assisted boundary proposals evaluated by annotation time, temporal accuracy, and agreement.

Acknowledgment: *This work was supported by the Student Scientific Research Projects of Electric Power University under Grant Nos. ĐTNH.55/2026 and ĐTNH.58/2026.*

REFERENCES

[1]. Vondrick et al., *"Anticipating Visual Representations from Unlabeled Video"*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016).

[2]. J. Carreira and A. Zisserman, *"Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset"*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2017).

[3]. W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, *"The Kinetics Human Action Video Dataset"*, arXiv preprint arXiv:1705.06950, (2017).

[4]. F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, *"ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding",* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 961–970, (2015).

[5]. M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, *"The Pascal Visual Object Classes Challenge: A Retrospective"*, International Journal of Computer Vision, Vol. 111, No. 1, pp. 98–136, (2015).

[6]. H.-N. Tran, H.-Q. Nguyen, H.-G. Doan, T.-H. Tran, T.-L. Le, and H. Vu, *"Pairwise-Covariance Multi-view Discriminant Analysis for Robust Cross-view Human Action Recognition"*, IEEE Access, Vol. 9, pp. 76097–76111, (2021).

[7]. Huong-Giang Doan, Thanh-Hai Tran, Hai Vu, Thi-Lan Le, Van-Toi Nguyen, Sang Viet Dinh, Thi-Oanh Nguyen, Thi-Thuy Nguyen, and Duy-Cuong Nguyen, *"Multi-view Discriminant Analysis for Dynamic Hand Gesture Recognition"*, Asian Conference on Pattern Recognition (ACPR), Vol. 1180, pp. 196–210, (2020).

[8]. A. Dutta and A. Zisserman, *"The VIA Annotation Software for Images, Audio and Video,"* Proceedings of the 27th ACM International Conference on Multimedia, (2019).

[9]. H. Zhao, A. Torralba, L. Torresani, and Z. Yan, *"HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization"*, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8668–8678, (2019).

[10]. CVAT.ai Corporation, *"Computer Vision Annotation Tool (CVAT)"*, Zenodo, (2023).

[11]. J. Cohen, *"A Coefficient of Agreement for Nominal Scales"*, Educational and Psychological Measurement, Vol. 20, pp. 37–46, (1960).

[12]. M. Tkachenko et al., *"Label Studio: Data Labeling Software"*, (2020–2025).

[13]. O. Crasborn and H. Sloetjes, *"Enhanced ELAN Functionality for Sign Language Corpora"*, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC), (2008).

[14]. R. Pontius et al., *"Death to Kappa: Birth of Quantity Disagreement and Allocation Disagreement for Accuracy Assessment"*, International Journal of Remote Sensing, Vol. 32, pp. 4407–4429, (2011).

[15]. G. Benitez-Garcia, J. Olivares-Mercado, G. Sanchez-Perez, and K. Yanai, *"IPN Hand: A Video Dataset and Benchmark for Real-Time Continuous Hand Gesture Recognition"*, Proceedings of the 25th International Conference on Pattern Recognition (ICPR), pp. 4340–4347, (2021).

TÓM TẮT

Công cụ gán nhãn video theo từng khung hình cho cử chỉ động và ứng dụng trong ngôn ngữ ký hiệu Việt Nam

Nhận dạng hành động con người trong video đóng vai trò quan trọng trong tương tác người–máy, đặc biệt đối với ngôn ngữ ký hiệu và điều khiển thiết bị thông minh. Nghiên cứu này đề xuất công cụ gán nhãn video ở mức khung hình cho ngôn ngữ ký hiệu tiếng Việt, hỗ trợ phân đoạn thời gian chính xác và xuất dữ liệu JSON phục vụ học sâu. Một bộ dữ liệu gồm 15 lớp cử chỉ động cũng được xây dựng bằng hệ thống thu thập đa góc nhìn. Chất lượng gán nhãn được đánh giá bằng Temporal IoU, Cohen's Kappa và thời gian gán nhãn, với kết quả đồng thuận lần lượt đạt 0.9470 và 0.9888.

Từ khóa: Nhận dạng hành động; Học sâu; Gán nhãn video; Temporal IoU; Cohen's Kappa; Phân đoạn theo thời gian; Đánh giá gán nhãn.