Journal of Scientific & Industrial Research
Vol. 82, April 2023, pp. 426-431
DOI: 10.56042/jsir.v82i04.72389

A Two-Stage Image Frame Extraction Model -ISLKE for Live Gesture Analysis on Indian Sign Language

Hyma J* & Rajamani P

CSE, GIT, GITAM University (Deemed to be), Gandhi Nagar, Rushikonda, Visakhapatnam, Andhra Pradesh 530 045

Received 17 May 2022; revised 21 September 2022; accepted 07 October 2022

The new industry revolution focused on Smart and interconnected technologies along with the Robotics and Artificial Intelligence, Machine Learning, Data analytics etc. on the real time data to produce the value-added products. The ways the goods are being produced are aligned with the people's life style which is witnessed in terms of wearable smart devices, digital assistants, self-driving cars etc. Over the last few years, an evident capturing of the true potential of Industry 4.0 in health service domain is also observed. In the same context, Sign Language Recognition- a breakthrough in the live video processing domain, helps the deaf and mute communities grab the attention of many researchers. From the research insights, it is clearly evident that precise extraction and interpretation of the gesture data along with an addressal of the prevailing limitations is a crucial task. This has driven the work to come out with a unique keyframe extraction model focusing on the preciseness of the interpretation. The proposed model ISLKE deals with a clustering-based two stage keyframe extraction process. It has experimented on daily usage vocabulary of Indian Sign Language (ISL) and attained an average accuracy of 96% in comparison to the ground-truth facts. It is also observed that with the two-stage approach, filtering of uninformative frames has reduced complexity and computational efforts. These key leads, help in the further development of commercial communication applications in order to reach the speech and hearing disorder communities.

Keywords: Classification, Clustering, Featured learning, Image processing, Region of interest, Video summarization

Introduction

New technologies have been evolved for efficient indexing and searching over widely available video content in digital platforms. One of the most popular emerging techniques is Video summarization.[1–3] It mainly concentrates on identifying salient features in the video content to get simplified, good-quality visual abstraction and thus to provide a short summary of the video sequence. This salient feature extraction is a significant phase in many video processing applications. In par with this, the recent research has resulted in promising developments in gesture-based applications.[4] To be specific, people with speech disabilities cannot express their ideas, feelings, or opinion as they can't speak. Sign language is the only medium of their communication which is in non-verbal form that comprises gestures. Gestures can be movement of a hand, face, or head to make them expressive which can be classified as simple and complex gestures, where a simple gesture is a single gesture and a complex gesture is a series of simple gestures.

*Author for Correspondence
E-mail: hjanapan@gitam.edu

Generally, a gestured video consists of a large number of frames, and all these frames do not carry significant information to identify a gesture. A very minimal number of frames is sufficient for it.[5,6] This can be ensured with effective keyframe detection models. Keyframe extraction process is one of the vital steps in many video processing based applications such as object detection, video annotation, automatic speech generation, video compressions and transmissions etc. The underlying idea in all these is to identify informative frames from the sequence of video frames. A frame can be informative or can be called a keyframe only when it best describes the summary of a scene.[6] As per the study, keyframe extraction methods are classified based on boundaries, perceptual content, motion, clustering, etc.[7] Among them, clustering-based keyframe extraction is more familiar and suitable, because of its adapted nature of diversified video formats. Identifying meaningful keyframes from a large number of frames is quite a challenging task in various video domains. However, very little experimentation has been noticed on keyframe extraction in sign language gesture detection.[8–10] Thus, the proposed work is aimed at achieving this

task with high efficiency using a two-stage approach. In accordance with this, the working system is planned to be organized into the following phases.

Problem Statement

Detecting and interpreting the movements of the human body is a major source of input for most of the gesture based Human Computer Interaction (HCI) applications. Sign language recognition is one such application and the primary focus of this work is to interpret these gestures. In view of this, the literature study witnessed various sign languages such as American sign language, Indian Sign Language (ISL), Chinese Sign Language, Turkish Sign Language. According to World Health Organization (WHO) statistics there is 6.3% of Hearing impairment and 7.5% of Speech impairment. This ignited us to focus on this research to address these disabilities. Observing India as a vast country with a fusion of various cultures and languages, English is one of the mostly used languages. The current ISL is based on English language and extension to regional languages is also initiated. Here the prominence of the ISL is viewed where the development scope of technological aids can be extended to all the regional languages for obtaining the common standard in addressing the difficulties of the community. In view of this, the proposed work has chosen ISL as a standard. The ISL vocabulary is designed in the form of gestured videos. To interpret the vocabulary by a computational model, it is very much necessary to decompose the video into frames which includes both informative and non-informative. To make the learning more flexible, the model has to identify unique keyframes out of redundant informative frames. Emphasizing this, the model aimed to summarize the ISL videos with short storyboard generation which in turn facilitates the automatic identification of the vocabulary.

Proposed Methodology

In video applications, keyframes are often used for summarizing the video content, authenticating the video, copyright protection, and so on. In the process of extracting the Keyframes from a video, frames containing maximum information are to be identified. This can be done by extracting the features from the image frames where each feature describes a different characteristic. Some of the general features are spatial, temporal, color characteristics, structural characteristics[7] such as texture, edges, curves, corners, histogram, interest points, etc. In some applications, it may be sufficient to extract only a single feature in the form of a feature vector, but it is not always the case. For example, multimodal processing requires working with different features in the form of feature space. In the literature, it has been observed that video summarization techniques have been categorized based on feature learning, clustering, event, shot selection[11], trajectory[12], etc. This work proceeds with the hybridized technique of feature and cluster-based learning with a major focus on sign language recognition - a promising field in gesture identification and analysis approach.

A higher-level abstraction of the proposed compact framework of gesture video summarization is organized at various stages namely - Preprocessing, Feature Learning, Candidate frameset generation, Unique Keyframe extraction, Storyboard representation and is depicted in Fig. 1. The input video is initially decomposed and preprocessed to obtain noise free, good quality frames, followed by a feature learning process to generate a candidate frameset. However, it is necessary to incorporate a unique keyframe extraction process to eliminate less informative frames from the candidate frameset that finally yields to a storyboard representation. The process is detailed as follows with the proposed algorithm ISLKE.

Candidate Frameset Generation

Generally, in the gesture identification process, picking up an accurate frameset having informative and non-informative gestured data is a key issue. Aiming in this aspect, the work aimed to develop an adaptive model that is well suited for the targeted application by the process of keyframe extraction. The entire process flow of the proposed model is organized in two stages:

Fig. 1 — Higher level abstraction

Candidate frameset generation and clustering the candidates to identify the selective frames required for short storyboard generation. Initially, after the frame decomposition, a selective threshold-based filtering approach is applied to effectively obtain redundant free candidates. A two-stage process is observed to be necessary, in order to filter the unnecessary frames at each stage. This leads to reduce the overburden of huge data in the later processing stage in case of long length videos.

In the stage of candidate frame set generation, canny multistage edge detection[13,14] filter is used to identify the features. Basing upon this feature vector, a global threshold G_t and adjacent frame difference AD is computed as given in Eq. 1 (a & b) as a part of structural similarity evaluation method[15], concentrates on highly structured components and strong correlation of images. This evaluation leads to a decision-making process to obtain a redundant free candidate frameset that majorly impacts the calculation speed.

$$AD_i = Abs\_Diff(f_i, f_{i+1}) \quad \ldots (1a)$$

$$G_t = Mean + (rand \times Std_{dev}) \quad \ldots (1b)$$

Unique Keyframe Extraction

From the observations, though the generated candidate frame set is redundant free, it is still huge in size. Concentrating on hand gesture recognition, region of interest (ROI)[16,17] is derived with the help of local maxima and minima[18] of nodal points[19,20] as shown in the Fig. 2. To facilitate more efficient storyboard retrieval, the process further proceeds with an ROI based clustering on the candidate frameset by using image hashing. The hash differences between adjacent frames are compared with an average cutoff threshold, and are assigned to clusters accordingly. The key pose clusters are formed with frames having high similarity and thus the center frame within each cluster can be treated as a keyframe. These are highly relevant and as a sequence replicate the summarized version of the original video.

Fig. 2 — Gesture nodal points

Algorithm: (ISLKE)

Input: Videos of ISL

Output: Frames of unique key poses

Step1: Frame decomposition, Let 'n' be the number of obtained frames

Step2: Redundant frame removal,

Step 2.1:- for i =1 to n ,

$AD_i = Abs\_Diff(f_i, f_{i+1})$

Step 2.2:- compute $G_t = Mean + (rand \times Std_{dev})$

where, $Mean = \sum_{i=1}^{n} AD_i$

$$Std_{dev} = \sqrt{\sum_{i=1}^{n}(AD_i - Mean)}/_{n}$$

Step 2.3:-for i =1 to n, assign frame i to candidate frameset CF when $Flag_i = 1$ where

$$Flag_i = \begin{cases} 1 \text{ if } AD_i > G_t \\ 0 \text{ if } AD_i < G_t \end{cases}$$

Step 3: - Frame Clustering

Step 3.1:- for i =1 to n , compute hash difference of nodal region

$HD_i = Abs\_Diff(h_i, h_{i+1})$

Step 3.2:- compute average cutoff threshold $T_h = \sum_{i=1}^{n} HD_i$

Step 3.3:- Let Cluster count, CC =1

for i =1 to n,

if $HD_i < T_h$

add frame i to Cluster CC ,

else

start a new cluster CC = CC+1

assign frame i to CC

Step 4:- Unique keyframe extraction-choose center frame from each cluster

Step5:- Summarize and generate short story board of ISL signs.

Results

The live gesture analysis is a complex task for the learning model, hence it has to be processed in such a way to reduce the uninformative content and only retain the essential features which would result in avoiding the tedious task involved. The proposed method is applied to ISL videos gathered from the Indian Sign Language portal. Initially, the method is aimed to process short videos of 3 to 5 seconds which express the daily usage vocabulary. To verify the validity of the data interpretation, it is cross verified with certified ISL Interpreters and also collected the ground truth images to further confirm the efficacy of the proposed model.

The model carries with various stages, and at the first stage of decomposition, for a duration of 3 to 5 secs with spatial resolutions of 1920 × 1080, 1080 × 720, 1280 × 980 has generated around 300 frames, which are further filtered with reduction phase and able to eliminate more than half of the uninformative frames as shown in Table 1, to attain a candidate frameset. Clustering is formed on hashes of the selected regions and unique frames are extracted from each cluster. From Fig. 3, it is witnessed that the model is able to retrieve the most appropriate keyframes using structural differences. The obtained sequence is consistent with the input video and so able to recapitulate.

The proposed two stage threshold based model succeeded in filtering out the uninformative frames in first stage and the redundant frames in second stage. By applying the clustering technique with an optimal threshold (identified by random experimentation) the model extracted unique frames in a proper sequence. These keypose clusters with an appropriate order is very much essential to identify the vocabulary of the video gestures.

It is understandable from the interpretation of the ISL videos that around half of the frames are the candidates that carries the factual information contributing towards the output. So, the proposed model aimed to filter the remaining unnecessary frames that actually reduce the video length in the further processing. The experimental evaluation is carried out on 150 ISL videos of daily usage vocabulary and the results of frame reduction for 6 sample videos are given in Table 1. An average reduction rate of around 46.92% simplified the further process with less computational cost.

The obtained reduced frameset or candidate frameset is given as input to the clustering process. Cluster analysis in Table 2 is presented with various

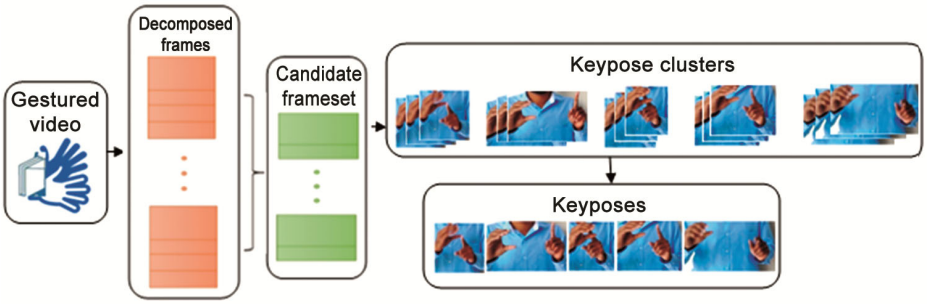

Fig. 3 — Frame extraction with sample ISL video

Table 1 — Data volume reduction

Video	Total frames	Reduced frameset	% of reduction
ISL sample video 1	166	92	55.42
ISL sample video 2	215	87	40.46
ISL sample video 3	144	77	53.47
ISL sample video 4	244	117	47.95
ISL sample video 5	183	78	42.62
ISL sample video 6	209	87	41.62
Average			46.92

Table 2 — Clustering analysis

Video	No of clusters	Cluster	Cluster size	Mean distance	Min distance	Max distance	Standard deviation	Intra cluster similarity purity percentage	Inter cluster similarity purity percentage
ISL sample video 1	7	c1	2	5	5	5	0	100%	98.75%
		c2	1	—	—	—	—	—	
		c3	10	2.8	1	5	1.1	97.18%	
		c4	3	2	2	2	0	98%	
		c5	14	1.84	0	4	1.459	96.25%	
		c6	5	2.25	0	4	1.78	97.12%	
		c7	2	4	4	4	0	98.98%	
ISL sample video 2	4	c1	5	1.75	0	5	1.92	97%	99.15%
		c2	4	2	1	3	0.81	97.01%	
		c3	18	2.23	0	5	1.307	96.78%	
		c4	1	—	—	—	—	—	
		c5	1	—	—	—	—	—	

Table 3 — Accuracy and error rates

Video	SSIM	Absolute error
ISL sample video 1	97.3	2.7
ISL sample video 2	98.2	1.8
ISL sample video 3	96.4	3.6
ISL sample video 4	95.7	4.3
ISL sample video 5	98.3	1.7
ISL sample video 6	96.2	3.8
Average	97.01	2.98

Table 4 — Computational time

	Video 1 (3 milli seconds)	Video 2 (10 milli seconds)
Preprocessing & feature Learning	2.82	4.49
Candidate frameset	38.95	68.48
Hand detection and ROI computation	54.89	175.73
Unique keyframe extraction	1.85	166.62

metrics like number, size, mean, standard deviation etc. For each input sample, the number of clusters and its size is obtained according to the key variations of the frame transitions. And are further validated with deviation, intra and inter cluster similarity measures. The keyframes obtained from the clusters when compared with the ground truth images have covered all the informative frames which are necessary to generate the short story board sequence. The applied granular level threshold filtering helped to reach the final set of ground truth images required to interpret the communicated ISL word labels. From the arrived results, it is witnessed that an average of more than 97% intra cluster similarity and more than 98% inter cluster similarity purity percentages are fair enough to proceed with the unique keyframe extraction from them.

The efficacy of the model is depicted with the comparative analysis between ground truth images and extracted keyframes in Table 3. The resulted structural similarity index proved that the model is able to extract all the appropriate keyframes with an average SSIM of 97.01%. The high score of more than 95% on all key poses indicates the accurate keyframe extraction of the ISL video. Phase-wise computational time is shown in Table 4, and it is observed that the relative effort is proportionally increasing with the length of the video.

Conclusions

A unique keyframe extraction technique ISLKE based on feature learning is proposed for sign language recognition. The proposed method succeeded in extracting the accurate key frames that are significant in conveying the information through gestures. Initially, the candidate frameset is obtained as an alternate frame sequence after the removal of redundancy. Finally, the optimal unique keyframe set is extracted from the clusters which are formed with the ROI-based structural analysis. The efficacy of the proposed model is validated against the ground truth facts and the experimental results have shown that it attained fair accuracy and accomplished in extracting the complete frameset required to generate the short storyboard of the ISL vocabulary. The systemized approach of the model has given visible results in terms of efficiency and computational complexity on videos of different lengths and resolutions. As a future work, an adaptive sequential deep learning model can be extended over the obtained keyframes to accurately recognize the sign language gestures.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is supported with funding through SEED grant No: Ref: F. No 2021/0026by GITAM University (Deemed to be).

References

1 Apostolidis E, Adamantidou E, Metsai A I, Mezaris V & Patras I, Video summarization using deep neural networks: a survey, *Proc IEEE*, **109(11)** 2021 1838–1863, doi: 10.1109/JPROC.2021.3117472.

2 Basavarajaiah M & Sharma P, Survey of compressed domain video summarization techniques, *ACM-Comput Surv*, **52(6)** (2020) doi: https://doi.org/10.1145/3355398.

3 Workie A, Sharma R & Chung Y K, Digital video summarization techniques: A survey, *Int J Eng Res Technol*, **09(01)** (2020) http://dx.doi.org/10.17577/ijertv9is010026.

4 Nama T & Deb S, Teleportation of human body kinematics for a tangible humanoid robot control, In *Cognitive Computing for Human-Robot Interaction* (Academic Press) 2021, 231–251, https://doi.org/10.1016/B978-0-323-85769-7.00011-2.

5 Schoeffmann K, Del Fabro M & Szkaliczki T, Keyframe extraction in endoscopic video, *Multimed Tools Appl*, **74** 2015) 11187–11206, https://doi.org/10.1007/s11042-014-2224-7.

6 Parihar A S, Mittal R, Jain P & Himanshu, Survey and comparison of video summarization techniques, *5th Int Conf Comput Commun Signal Process* (Chennai, India) 2021, 268–272, doi: 10.1109/ICCCSP52374.2021.9465347.

7 Shi Y, Yang H, Gong M, Liu X & Xia Y, A fast and robust key frame extraction method for video copyright protection, *J Electr Comput Eng*, (2017) 1–7, https://doi.org/10.1155/2017/1231794.

8 Huang S, Mao C, Tao J & Ye Z, A novel Chinese sign language recognition method based on keyframe-centered

clips, *IEEE Signal Processing Letters*, **25(3)** (2018) 442–446, doi: 10.1109/LSP.2018.2797228.

9 Pan W, Zhang X & Ye Z, Attention-based sign language recognition network utilizing keyframe sampling and skeletal features, *IEEE Access*, **8** (2020) 215592–215602, doi: 10.1109/ACCESS.2020.3041115.

10 Yan Y, Li Z, Tao Q, Liu C & Zhang R, Research on dynamic sign language algorithm based on sign language trajectory and key frame extraction, *IEEE 2nd Int Conf Electron Technol (ICET)* (Chengdu, China) 2019, 509–514, doi: 10.1109/ELTECH.2019.8839587.

11 Elahi G M M E & Yang Y-H, Online learnable keyframe extraction in videos and its application with semantic word vector in action recognition, *Pattern Recognit*, **122** (2022) https://doi.org/10.1016/j.patcog.2021.108273.

12 Haq H B, Asif M & Ahmad M B, Video summarization techniques: a review, *Int J Sci Technol Res*, **9(11)** (2020) 146–153.

13 L. Xuan & Z. Hong, An improved canny edge detection algorithm (IEEE) 2017, 275–278, doi: 10.1109/ICSESS.2017.8342913.

14 Song R, Zhang Z & Liu H, Edge connection based canny edge detection algorithm, *Pattern Recognit Image Anal*, **27** (2017) 740–747, https://doi.org/10.1134/S1054661817040162.

15 Hu H, Wang W, Zhou W, Zhao W & Li H, Model-aware gesture-to-gesture translation, *IEEE/CVF Conf Comput Vis Patt Recognit* (Nashville, TN, USA) 2021, 16423-16432, doi: 10.1109/CVPR46437.2021.01616.

16 Gupta A, Mohatta S, Maurya J, Perla R, Hebbalaguppe R & Hassan E, Hand gesture based region marking for tele-support using wearables, *IEEE Conf Comput Vis Pattern Recogn Work* (Honolulu, HI, USA) 2017, 386–392, doi: 10.1109/CVPRW.2017.53.

17 Bakheet S & Al-Hamadi, Robust hand gesture recognition using multiple shape-oriented visual cues, *EURASIP J Image Video Process*, **26** (2021) 1–18, https://doi.org/10.1186/s13640-021-00567-1.

18 Lee D L & You W S, Recognition of complex static hand gestures by using the wristband-based contour features, *IET Image Process*, **12(1)** (2018) 80–87, doi: 10.1049/iet-ipr.2016.1139www.ietdl.org.

19 Ponnarassery P K, Agnihotram G & Naik P, Human pose simulation and detection in real time using video streaming data, *S N Comput Sci 1*, **148** (2020), https://doi.org/10.1007/s42979-020-00153-8.

20 Awasthi V & Agnihotri A, A structural support vector machine approach for biometric recognition, *Int J Comput Info Eng*, **15(4)** (2021) 273–280.