Assistive technology for relieving communication lumber between hearing/speech impaired and hearing people

Rini Akmeliawati[1], *Donald Bailey*[1,2], *Sara Bilal*[1], *Serge Demidenko*[2,3], *Nuwan Gamage*[4], *Shujjat Khan*[2], *Ye Chow Kuang*[4], *Melanie Ooi*[4], *Gourab Sen Gupta*[2]

[1]*Faculty (Kulliyyah) of Engineering, International Islamic University Malaysia, Jl, Gombak 53100, Kuala Lumpur, Malaysia*
[2]*School of Engineering and Advanced Technology, Massey University, New Zealand, Private Bag 11222, Palmerston North 4442, New Zealand*
[3]*Centre of Technology, RMIT University Vietnam, 702 Nguyen Van Linh Blvd, Ho Chi Minh City, HCMC, Vietnam*
[4]*School of Engineering, Monash University Malaysia, Jl Lagoon Selatan, 46150, Selangor Darul Ehsan, Malaysia*
E-mail: serge.demidenko@rmit.edu.vn

Published in *The Journal of Engineering*; Received on 4th February 2014; Accepted on 15th May 2014

Abstract: This study proposes an automatic sign language translator, which is developed as assistive technology to help the hearing/speech impaired communities to communicate with the rest of the world. The system architecture, which includes feature extraction and recognition stages is described in detail. The signs are classified into two types: static and dynamic. Various types of sign features are presented and analysed. Recognition stage considers the hidden Markov model and segmentation signature. Real-time implementation of the system with the use of Windows7 and LINUX Fedora 16 operating systems with VMware workstation is presented in detail. The system has been successfully tested on Malaysian sign language.

1 Overview

'Sign language' (SL) is a highly structured non-verbal language utilising both manual and non-manual communications. Manual communication consists of movements and orientation of hand/arm conveying symbolic meaning, whereas non-manual communication involves mainly facial expression (as shown in Fig. 1*a*), head movement, body posture and orientation, which help in augmenting the meaning of the manual signs. Furthermore SL consists of static signs, which are mainly the alphabets (shown in Fig. 1*b*), and dynamic signs, which involve some motions as shown in Fig. 1*c*.

SLs have a systematic and complex structure of grammar that consists of isolated signs (words) and continuous signs (sentences) differing from one country to another.

Everyday communication with the hearing population poses a major challenge to those with hearing loss. For many people who were either born with hearing impairment or became impaired later in their lives, SL is used as their main language. Spoken languages such as English, Malay and others are often learnt only as a second language. As a result, their reading and writing skills are often below average as they mostly opt to converse in SL. Although some can read, many others fail in cases where reading is needed.

Common current options for alternative communication modes include cochlear implants, writing and interpreters. Cochlear implants are small and complex electronic devices that can help to provide a sense of sound to a person who is profoundly hearing/speech impaired or severely hard-of-hearing. The use of a cochlear implant requires both a surgical procedure and significant therapy to learn or re-learn the sense of hearing. Not everyone acquires learning at the same level with this device and the device is relatively expensive. Handwriting is another alternative for communicating with hearing/speech impaired people. However, most of the hearing/speech impaired people cannot communicate well through written language because they use SL as their preferable language for communicating. Interpreters are commonly used within the hearing/speech impaired community, but interpreters can charge high hourly rates and be awkward in situations where privacy is of high concern, such as at a doctor or lawyer's office. In addition, the number of interpreters is very limited particularly in developing countries like Malaysia [1].

'Automatic SL translator' (ASLT) is an automated system using advanced technology to translate a particular SL into a readable language such as English, Malay, Chinese etc. The existing ASLT systems generally use the following:

i. *DataGlove* or *CyberGlove* [2]: A specially built electronic glove worn by a signer. The glove has built-in sensors, which detect and transmit information on the hand posture as illustrated in Fig. 2.

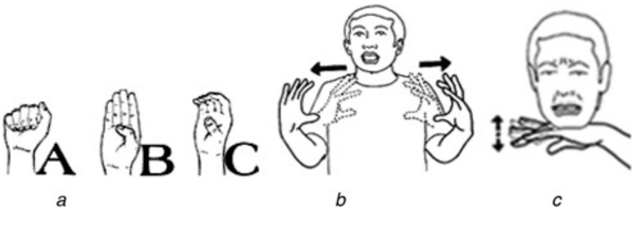


Fig. 1 *Overview*
a Alphabet letters '*A*', '*B*' and '*C*'
b Dynamic sign using two hands 'big*'*
c Facial expression in sign 'dirty*'* [1]

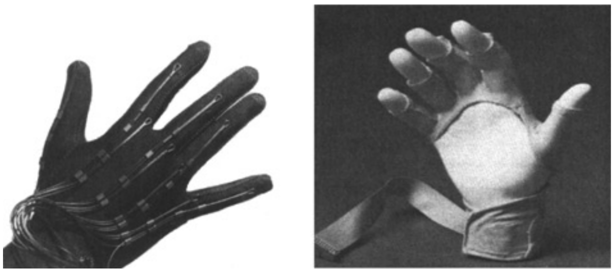


Fig. 2 *DataGlove (left) and CyberGlove (right) [2]*

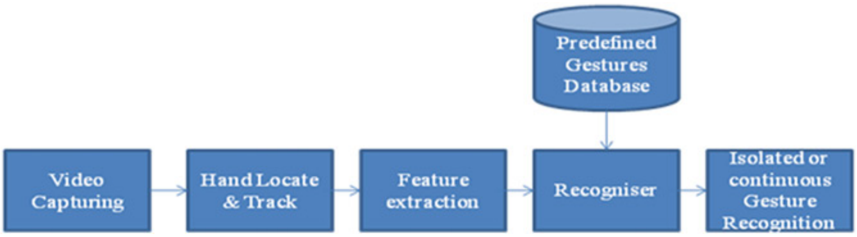

Predefined Gestures Database
Video Capturing
Hand Locate & Track
Feature extraction
Recogniser
Isolated or continuous Gesture Recognition

Fig. 3 *Structure of a complete ASLT*

ii. *Vision-based approaches:* where a camera is used to capture images of a person who is signing by using either a coloured glove or just bare hands. The major advantage of this approach compared with the application of the DataGlove or CyberGlove is the flexibility. It can be developed to include non-manual signs such as recognition of facial expressions and head movements as well as lip-reading. In addition, the position of the signer's hand with respect to other body parts could be identified by using the vision system. Owing to the above-mentioned advantages, this paper focuses on the vision-based ASLT.

This paper is organised as follows. Section 2 presents the system architecture, starting from the video image acquisition, hand location and tracking and feature extraction. Section 3 describes the recognition stage in which 'hidden Markov model' (HMM) and segmentation signature are detailed. Section 4 discusses the real-time implementation of the system. Section 5 discusses the performance of the system. Finally, this paper is concluded in Section 6.

2 System architecture

In general, the structure of vision-based automatic SL consists of four crucial stages; database collection, blob detection and tracking, feature extraction and the recognition stage.

The structure is shown in Fig. 3.

2.1 SL database collection

Most SLs databases are available only for educational and learning purposes. The SL database is required to build algorithms, expressing the nature of signs and covering the possibility of signing. Recent researches in SL recognition are conducted for SLs used in different countries.

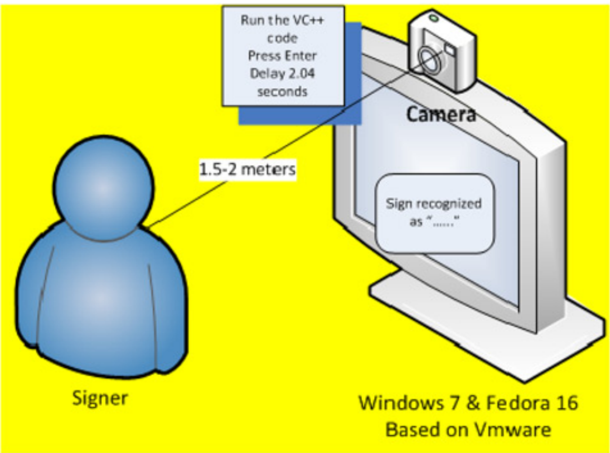


Fig. 4 *Sign language database collection system set-up*

The 'Malaysian SL' (MSL) database was collected by recording video of signers. The signer has to stand about 1.5–2 m in front of the camera, where his/her upper body is visible in the scene as shown in Fig. 4. Once the signer starts signing, the recording process is initiated. The video is then stored into the computer for further processing.

2.2 Blob detection and tracking

The second stage of the ASLT includes finding the face and two hand blobs in each video frame. There are several existing methods to accomplish these tasks such as skin colour approaches [3], motion constraint approaches [4] and the static background inclusion approaches [5]. However, it is often not reliable to model a skin colour where high variations of skin colours and different lighting conditions are present. Limitations arise from the fact that human skin can be defined in various colour spaces after applying colour normalisation. Therefore the model has to accept a wide range of colours, making it more susceptible to the noise. On the other hand, the motion systems assume that the hand is the fastest moving object in the image frame. This is not the case when the hand gesture does not carry a fast motion or the head gesture is stronger than the hand, that is, the head gesture is more active than that of the hand. In addition, the background inclusion approaches assume a static background. These approaches have many limitations in terms of reliability. Therefore those methods have drawbacks making them to be not a good choice for the SL recognition developed system. Therefore a novel hybrid system combining the appearance-based method and skin colour segmentation has been introduced:

(a) *Blob detection with Haar-like features:* Initialising the system with a face or specific hand shape will provide an adequate search region for skin pixels. 'Artificial neural network'-based methods, 'support vector machines' (SVMs) and other kernel methods have been used for face or hand posture detection. However, most of these algorithms use raw pixel values as features. This makes such systems sensitive to noise and changes in illumination. Instead, other approaches such as Haar-like features proposed by Viola and Jones [6] have been used in this paper to detect the face or hand as an initial stage. Meanwhile, the system has been initiated by detecting the face or hand region. The dimension of the colour space is not a big concern because the range of the skin pixels within the detected area has been obtained. This reduces the required memory space and processing time. To perform the skin detection from an image, the image needs to be converted from red–green–blue (RGB) to YC_bC_r colour space after the face or hand has been detected. Then, in order to find skin pixels that fall within the same colour space range of the detected hand or face, 10 × 10 pixels box from the centre of the face is extracted as shown in Fig. 5*c*. Finally, a range of skin pixels is specified based on the detected face, the distribution of skin pixels values is highlighted using C_b and C_r components and ($C_b - C_r$ or $C_r - C_b$) as an additional threshold while the luminance Y component is discarded. The segmented image is shown in Fig. 5*d*.

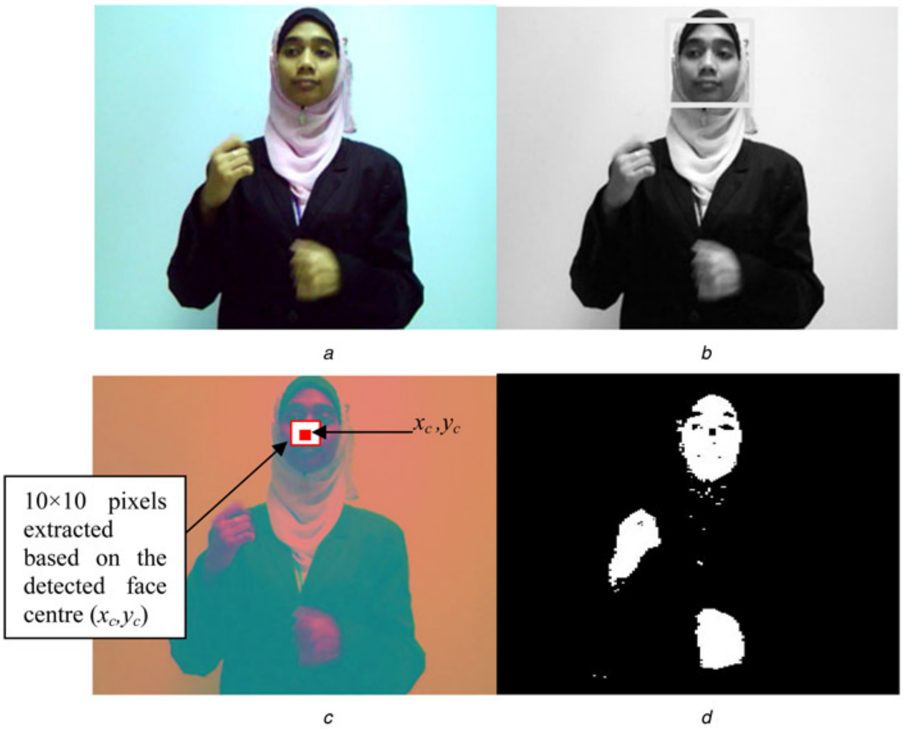

a
b
x_c, y_c
10×10 pixels extracted based on the detected face centre (x_c, y_c)
c
d

Fig. 5 *Blob detection with Haar-like features*
a Original RGB image
b Detected face on grey-scale image
c YC_bC_r image
d Segmented image based on C_b, C_r and $C_r - C_b$ threshold [7]

(b) *Blob labelling:* The detected and extracted face and two hand blobs must be identified and labelled. In this paper, the method developed in [8] that simultaneously labels contours and components in binary images is used. This method labels the blobs and identifies each blob based on certain criteria such as the size and perimeter. However in the developed system, the blob labelling has been achieved by applying the y-axis labelling approach as shown in Fig. 6, rather than using the size and the perimeter as mentioned in [8].

2.3 Feature extraction

Performance of any SL recognition system significantly depends on obtaining efficient features to represent pattern characteristics. There is no algorithm, which shows how to select the representation or to choose the features. The selection of features depends on the application [9]. There are many different methods to represent two-dimensional (2D) images such as boundary, topological, shape grammar and description of similarity. Features should be chosen so that they are intensive to noise-like pattern variations and the number of features should be small for easy computation. The hand posture shape feature, motion trajectory feature and hand position with respect to other human upper body (HUB) parts play an important role within the preparation stage of the gesture before SL recognition stage. There are static and dynamic signs in SL as mentioned earlier. The static signs usually come under the alphabet signs. This research has conducted tests on static signs as well as dynamic ones.

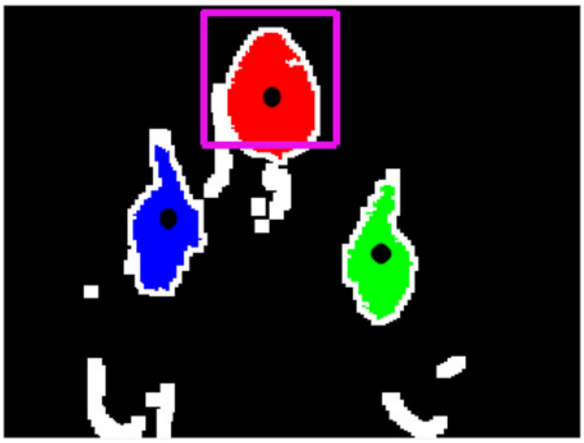


Fig. 6 *Blob labelling using different colours*

2.3.1 1D profile for finger detection in static signs: The boundary of a hand is extracted by subtracting the filled image from the eroded one. An important issue is to determine the internal profile of a detected hand boundary, which emphasises on the detection of fingers. This is done by allocating the centre of the target (x_c, y_c).

Then the radial distance $d_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}$ from the centroid (x_c, y_c) to each array point (x_i, y_i) is computed. After smoothing d_i for noise reduction using Savitzky–Golay smoothing filters, the local maxima in the 1D curve has been determined. The number of the peaks in the vector above the centre range determines the fingers, whereas the approximate distance between the peaks identifies the finger which is opened. This procedure is shown in Fig. 7.

2.3.2 High-level features for dynamic signs: This research investigated three types of features, which are: ‘geometric’ (shape), ‘motion’ and ‘location’.

Geometric features: Contours or edges are features, which can be used in any model-based technique as well as in non-model ones. The aim is to obtain similar values of features for similar hand

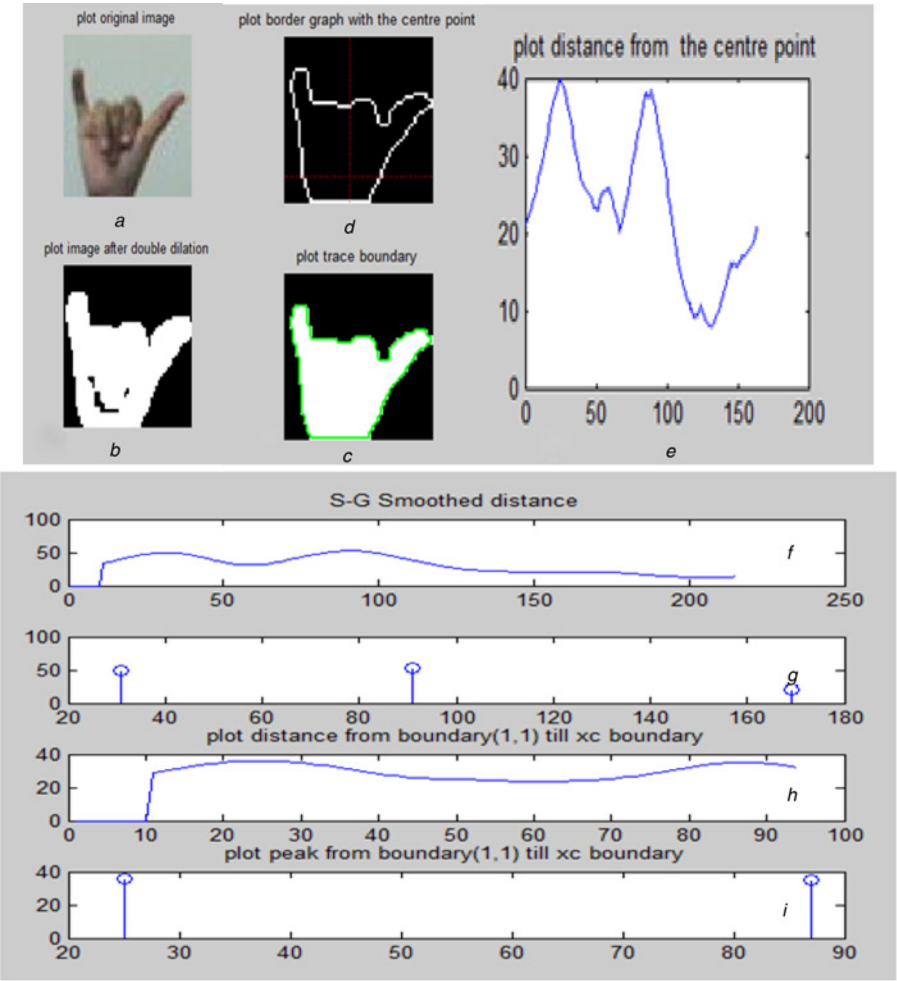


Fig. 7 *1D profile for finger detection in static signs*
a Detected hand
b Double dilation
c Filled image
d Boundary
e 1D profile
f Smoothed profile
g Local maxima
h and *i* Upper hand profile and maxima from its centre [10]

shapes and distant values for different shapes. This is done by choosing seven features (blob area, semi-major and semi-minor axes, orientation angle, compactness, roundness and eccentricity). These features are moment-based features from the grey-scale image. The definition of moments can be found in [11].

Motion features: To capture the dynamic characteristics of the hand posture, sequence of frames must be available with the hand motion. The main issue is to track the hand over the video frames and extract the centre point (x_c, y_c) from the hand posture over a sequence of successive frames. Following that, eight trajectory-based features are extracted from the detected blob. These eight features are stored in a feature vector. An n-dimensional feature vector represents numerical features of an object. These features are the centre point of the hand (x_c, y_c), difference between the consecutive centre points, average sum of x_c and y_c, velocity and angle.

Location features: In SL recognition systems, it is very important to find the hand location with respect to the head, the shoulder and the chest as this carries a lot of meaning. The HUB parts can be used as reference for static and dynamic signs. Various existing methods using geometric modelling, boosted classifiers and SVM have been introduced for HUB detection. However, real-time applications such as automatic SL recognition systems require a fast method for HUB part detection and tracking. Therefore, in this paper, a fast and robust search algorithm for HUB parts based on the study [12] has been introduced. It assumes that all body parts can be measured with respect to a head size. Initialising the system with finding a face provides an adequate search region for other HUB parts. The proposed system used Haar-like features and ‘AdaBoost’ algorithm for face region detection. Following that, other HUB regions are found based on a human figure adjusted for artists based on accurate 8-head size as shown in Fig. 8.

Tracking HUB region: Pose detection (or initialisation) is typically performed on the initial video frame followed by pose tracking where the pose parameters obtained from the current frame are used as the starting value for the subsequent video frame. Many methods exist to track pose using motion histories and optical flow technique. Other methods, such as continuously adaptive mean shift (CAMSHIFT) were designed for face and coloured object tracking using probability distribution [13]. CAMSHIFT is very sensitive to the change of the face colour over time. Therefore the tracking in this paper is based on the face centre point which is obtained from the skin binary image. This method of tracking performs better than the CAMSHIFT tracking method as shown in Figs. 9 and 10.

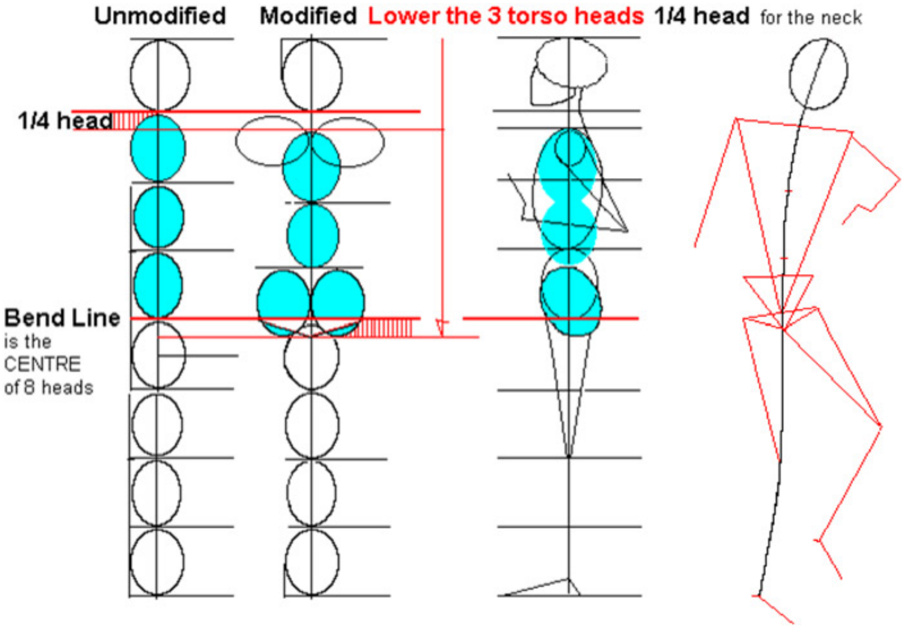


Fig. 8 *Accurate 8-head-high adult male human figure [12]*

3 Recognition of SL using HMM

In this research HMM has been chosen for training and testing the MSL translator system based on its ability to distinguish transitions between signs. It is sufficient to model a simple sign using a three state HMM. Meanwhile for more complex signs in terms of motion and sign duration more states are necessary [15]. Therefore in this paper a four state HMM have been used to model isolated signs, with no skip state. Training of the HMM-based recogniser was done using tools developed employing C and C++ programming languages under 'LINUX' operation system.

3.1 Isolated and continuous sign training, testing and recognition

Signs have been collected with at least ten repetitions for each sign. Then, signs passed through blob detection to feature extraction stages. After extraction of motion and geometric features from isolated and continuous signs, these features are concatenated in one text file. The most crucial issue during training and recognising isolated and continuous signs is the HMM topology, grammar structure and master label file (MLF) for words and sentence structure.

3.2 Grammar for isolated and continuous signs

3.2.1 Grammar for isolated signs: An example of the grammar file structure, which contains five isolated words 'car', 'chicken', 'dog', 'noodles' and 'school' is as follows

$$\$\text{gesture} = \text{car}|\text{chicken}|\text{dog}|\text{noodles}|\text{school}(\$\text{gesture})$$

3.2.2 Grammar for continuous signs: The grammar for sentences was chosen after analysing different grammar structures. Many different types of grammar were tested and the below sentence grammar structure was chosen

$$\text{pronoun}/\text{noun} + \text{verb} + \text{adjective} + \text{noun}$$

The grammar with both 'Start' sign and 'End' sign was tested. The start and end delay times are very crucial because it shows the beginning and ending of the sentence. In the developed system, it is not required for a signer to sign under certain restrictions, such as fixing a start sign. Therefore the start and end of the sign is taken as a random observation.

3.3 Recognition accuracy of isolated signs

The isolated signs comprised 42 signs from MSL database, which were collected randomly within the categories of pronoun, noun, verbs and adjective. The evaluation of different isolated sign combinations has been done by running different experiments in order to analyse the effect of feature vectors on the recognition accuracy.

3.3.1 Effect of features: example of four signs: Four signs 'green', 'yellow', 'water' and 'rice' have been chosen based on the similarities between each pair ('green' and 'yellow', 'water' and 'rice') to run the second set of experiments. Signs 'green' and 'yellow' are very similar to each other as illustrated in Figs. 11*a* and *b*. It is difficult even for human eyes to differentiate between these two signs.

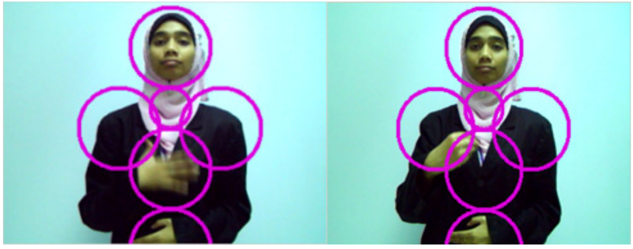


Fig. 9 *Tracking of HUB within two frames of sign 'white' using the proposed method [14]*

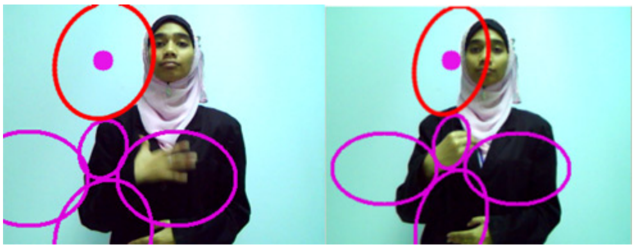


Fig. 10 *Tracking of HUB within two frames of sign 'white' using CAMSHIFT [13]*

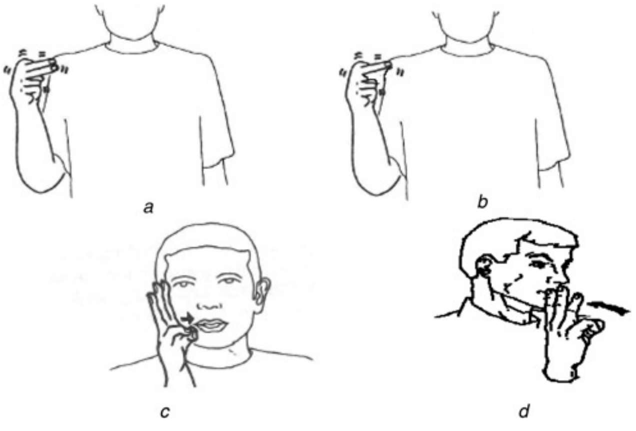


Fig. 11 *Signs*
a ‘*Green*’
b ‘*Yellow*’
c ‘*Rice*’
d ‘*Water*’ from MSL database [1]

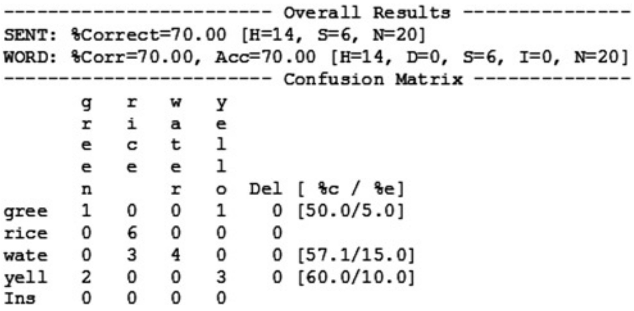
From Fig. 12, it is clear that the confusion mainly happened between signs that have similarities such as ‘green’ and ‘yellow’ because the motion trajectory is relatively the same. In such cases, one can estimate that geometric features could play a role in the accuracy enhancement. In fact, Fig. 13 shows that the existence of geometric features could drop the system accuracy. These could be verified as follows:

i. Signs ‘water’ and ‘rice’ have better accuracies in Fig. 13 than the confusion that happened between the two signs in Fig. 12. This is because of nature of the two signs; that is, slow dynamic motion (see Fig. 11).
ii. Signs ‘green’ and ‘yellow’ are very similar even for human vision (see Figs. 11*a* and *b*). Therefore none of the features could enhance the system accuracy in that case.

3.3.2 Effect of feature vectors: combining one- and two-hand signs: Another two sets of experiments have been launched while combining signs that use two hands and signs that use only one hand. The set of experiments employing eight feature vectors from one- and two-hand signs have been used. Although using eight feature vectors for one- and two-hand signs separately, the system accuracy was 75.9883 and 85.61%, respectively. However, if the two sets (for one- and two-hand signs) are combined (11 signs in total) the accuracy was 83.63%.

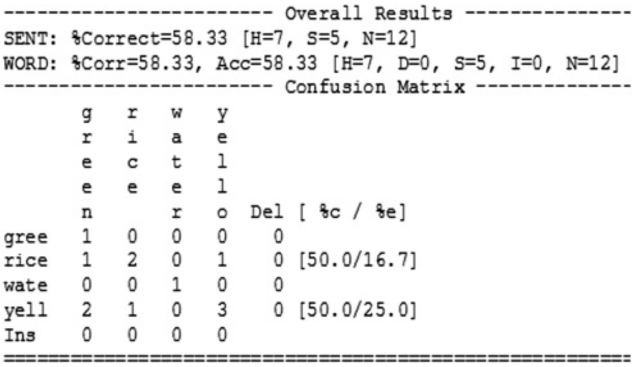
3.4 Recognition accuracy for continuous signs

Sentences with the grammatical structure ‘noun/pronoun, verb, adjective, noun’ were chosen for the recognition. In total 20 nouns, 5 pronouns, 5 adjectives and 7 verbs with a total lexicon of 37 words were employed.



```
------------------------ Overall Results ---------------
SENT: %Correct=70.00 [H=14, S=6, N=20]
WORD: %Corr=70.00, Acc=70.00 [H=14, D=0, S=6, I=0, N=20]
------------------------ Confusion Matrix --------------
       g   r   w   y
       r   i   a   e
       e   c   t   l
       e   e   e   l
       n       r   o  Del [ %c / %e]
gree   1   0   0   1    0 [50.0/5.0]
rice   0   6   0   0    0
wate   0   3   4   0    0 [57.1/15.0]
yell   2   0   0   3    0 [60.0/10.0]
Ins    0   0   0   0
```

Fig. 12 *Training and test results for ‘yellow’, ‘green’, ‘water’ and ‘rice’ signs (motion features)*



```
------------------------ Overall Results ---------------
SENT: %Correct=58.33 [H=7, S=5, N=12]
WORD: %Corr=58.33, Acc=58.33 [H=7, D=0, S=5, I=0, N=12]
------------------------ Confusion Matrix --------------
       g   r   w   y
       r   i   a   e
       e   c   t   l
       e   e   e   l
       n       r   o  Del [ %c / %e]
gree   1   0   0   0    0
rice   1   2   0   1    0 [50.0/16.7]
wate   0   0   1   0    0
yell   2   1   0   3    0 [50.0/25.0]
Ins    0   0   0   0
=======================================================
```

Fig. 13 *Training and test results for ‘yellow’, ‘green’, ‘water’ and ‘rice’ signs (motion and geometric features)*

In general, the duration of each sentence was between 2 and 3 s. Two ‘native’ signers signed the sentences in a natural way as they communicate with each other in the ‘Malaysian Federation of the Deaf’ society. Therefore no specific pauses between the signs within the sentences took place. The signed sentences from each signer were divided into two sets. The system was trained using 90% of the collected sentences and 10% of them were used for validating the system. HMM with four states without any states skipped was developed for training the sentences.

Sets of experiments were utilised using two signers. ‘Signer 1’ signed 172 sentences and ‘Signer 2’ signed 202 sentences. The experiments have been repeated 50 times with different datasets for training and testing from the 172 and 202 sentences, separately. First, eight motion features have been used to test the system accuracy. The recognition accuracies of sentences were 55.02 and 55.52%. Then the geometric and motion features were combined together to train and test the system. Using 172 and 202 sentences signed by two signers, the achieved recognition accuracies were 39.39 and 42%, respectively.

Published studies suggest that the recognition performance of the isolated gesture approaches over a natural discourse is deteriorated by the dynamics of an uninterrupted communication [16–23]. These dynamic constraints are subjective and related to the language experience as well as word choice skills of the signer. Just like a native speaker, a native signer not only signs fluently, but also utilises a broader vocabulary as compared with inexperienced ones (non-natives). Similarly, sometimes an experienced person may exploit the parallel nature of SL by conveying multiple ideas employing compound gestures through the process called modulation [24]. Another challenging aspect of continuous recognition is so-called co-articulation, which is a process of smoothly connecting two lexically different glosses [25]. Since the joining glosses are mostly distinct in their lexical components, their co-articulation may cause significant transformation in a parameter space disturbing their spatial, morphological or temporal components. For example, referring to frames in Figs. 14*b*–*d* we can observe the preparatory transitions after the end of a noun ‘I’ to another sign ‘Go’ in Figs. 14*a* and *e*, respectively. These lexically insignificant transitions in shape and location are hard to distinguish as invalid signs and consequently they may be mistranslated. Without employing any proper segmentation or synchronisation, severe context modification is eminent because the intermittent sign in Fig. 14*b* can be recognised as a verb ‘to cost’ or noun ‘money’.

A simple solution used by most of the existing systems is to introduce a synchronisation mechanism through a variety of *ad hoc* measures like pseudo pauses [26–28]. They are like the silent period between two words in a speech. These pauses provide the synchronisation, which makes the isolated gesture recognition system work

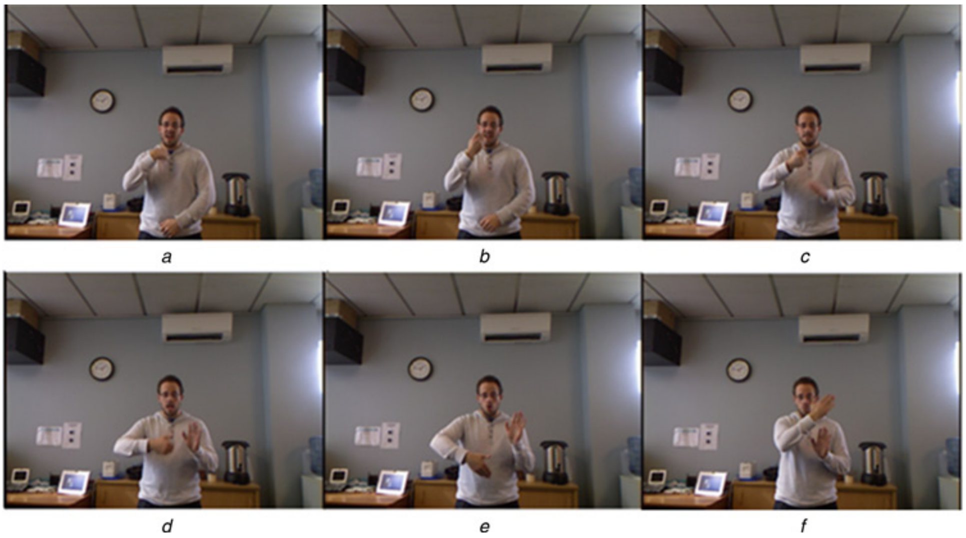


Fig. 14 *Recognition accuracy for continuous signs*
a and *e* Noun 'I' and verb 'Go' shown in frames
b–d Connected through Co-articulation frames

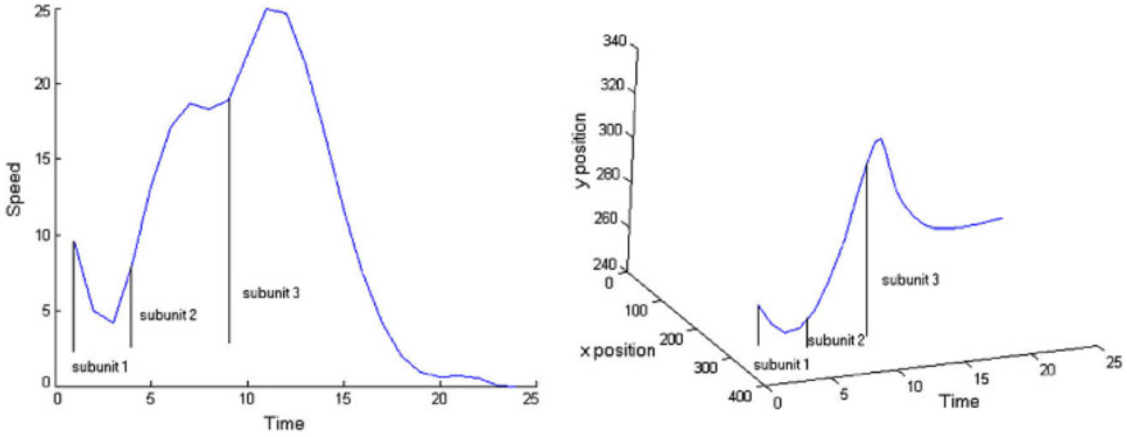


Fig. 15 *Subunit segmentation through directional variation [32]*

on a continuous discourse. The inter-sign pauses are explicitly inserted in many ways either by the exaggerated hold of every sign on its completion, bringing articulators back to a specific neutral position or by taking them out of the field of view. In another approach, they are explicitly triggered by some external means like signer must press a button/paddle after each sign by his/her toe. Obviously, these *ad hoc* segmentation measures simplify the recognition process by turning a continuous discourse into a co-articulation-free sequence of disjoint postures. Nevertheless, for a practical recognition system, these schemes are not too attractive because they disturb the natural prosody of a signer.

HMM-based approaches assume the co-articulation as a temporal variation. In such methods, each gesture in the vocabulary is modelled by an HMM [29, 30], which is a generative model based on likelihood and priors learned during the training phase. These approaches are reported to be robust and able to normalise any temporal inconsistencies. However, they require a huge amount of training data to obtain a system fully trained on a medium-sized vocabulary [16, 29]. These techniques are very useful in continuous speech recognition applications because of the phonological decomposition of a word into its basic units called phonemes. Now instead of training an HMM for each word, it is trained on its phonemes which numbers about 50–60, far lesser than the entire vocabulary. Unfortunately because of unavailability of any valid subunit of a sign or the very large amount of assumed subunits called 'cherems' (about 2500–3000), the HMM-based approaches have limited use. Alternatively, in the deterministic approaches also called 'direct segmentation methods', all valid signs in a

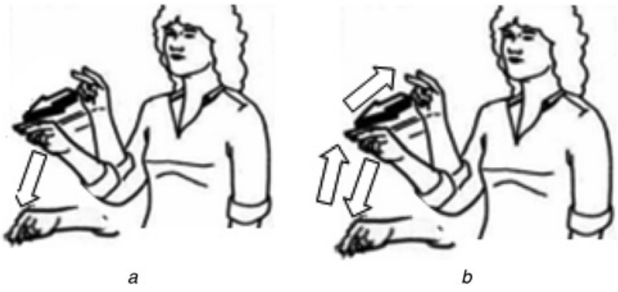


Fig. 16 *Verb sign 'to ask' is gesticulated by a single movement shown by arrows*
a Gesture for '*ask*'
b '*Asking*'

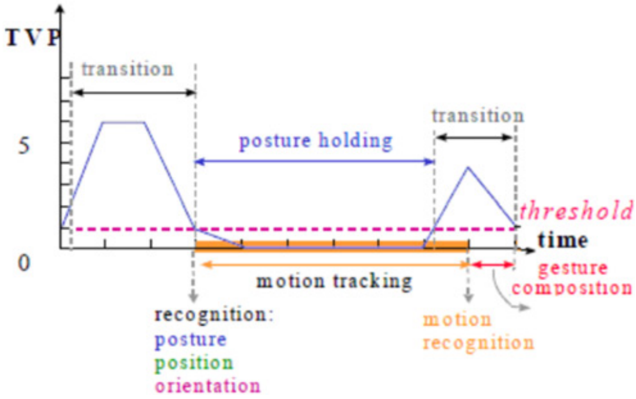


Fig. 17 *Time-varying parameter-based segmentation [33]*

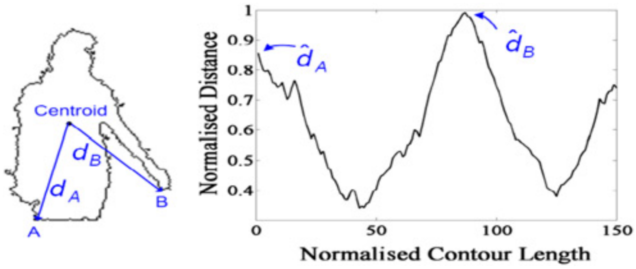


Fig. 18 *Motion signature formation [23]*

continuous sentence are extracted by detecting the sign boundaries with the help of different spatiotemporal features. Research [31] proposed a subunit extraction system based on the deterministic motion features like articulator's velocity. As shown in Fig. 15, it hypothesised the significant directional variation of an articulator as the main segmentation features of the subunits (cherems) of a sign.

Movement component of a gesture is the most significant part of a continuous discourse, which is considered for the segmentation. A majority of the existing models utilise the movement trajectories (2D or 3D) and their temporal derivatives (velocity and acceleration) as their segmentation features. Inspired from the pause-based

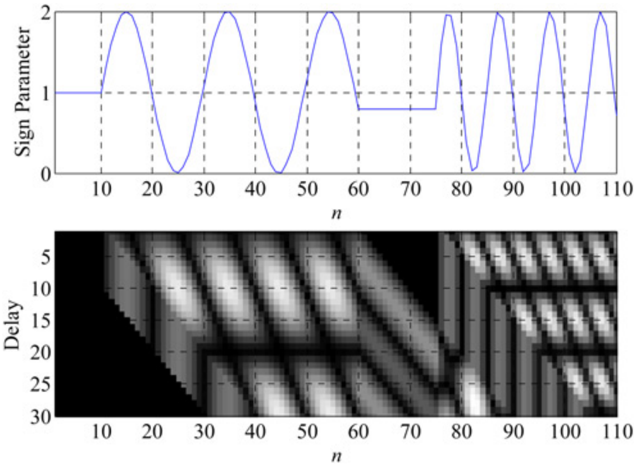


Fig. 20 *DAD signature of time-varying sign parameters*

speech segmentation, these schemes mainly rely on the energy of a continuous signal like local minima or the discontinuities.

The word segmentation of a natural SL discourse by a native signer results in high false positive rate because of unclear 'pauses' in the hand movement. The accuracy of most existing approaches deteriorates without imposing artificial pauses or exaggeration in the normal signing. Apart from the motion information of a gesture, there are few other unaddressed spatiotemporal cues to detect the sign boundaries. Some of these include a sudden change in articulator's direction, change in non-manual signs and short-termed repetitions. Sign repetition is frequently used in natural signing for a variety of signs like interrogative, explanative and indicative gestures. More importantly, the temporal references of a gesture are modified through the repetitions. For example, a verb sign 'to ask' is gesticulated by a single movement shown by arrows in Fig. 16*a*, but the repetition of the same sign (Fig. 16*b*) turns it into a present participle form 'asking'. Obviously, the sign repetition becomes a clear indication of the gesture boundary.

Research [33] proposes a deterministic segmentation scheme in which a word boundary is decided by observing the state of its time-varying parameters. As shown in Fig. 17, if the number of time-varying parameters drops below a specific threshold, the articulator is considered to be in a quasi-stationary state. Therefore the corresponding frame is taken as the end of the previous sign and all

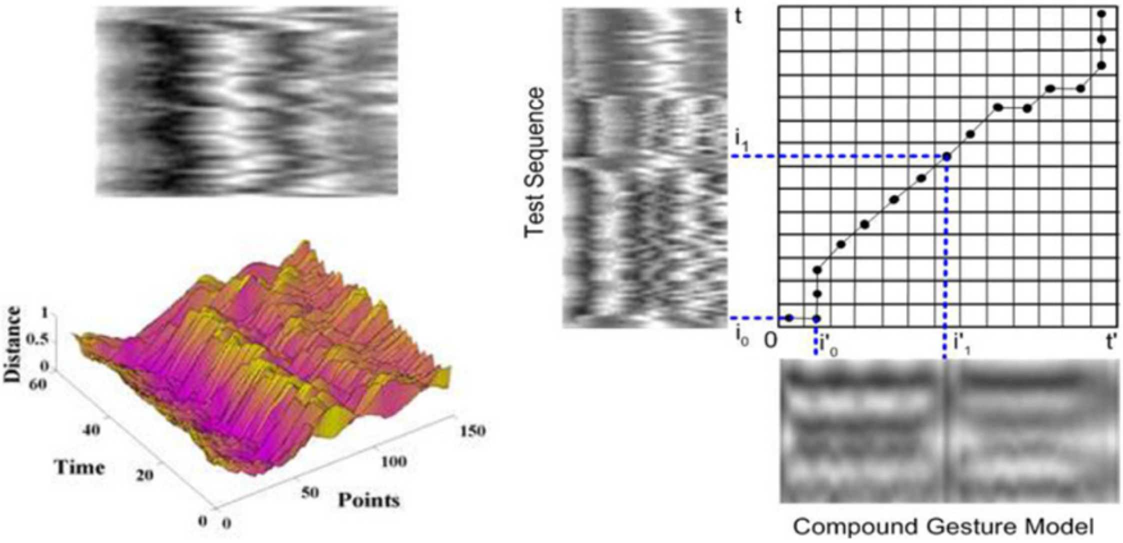


Fig. 19 *Left: signature modelling (2D and 3D); right: matching using dynamic time warping [23]*

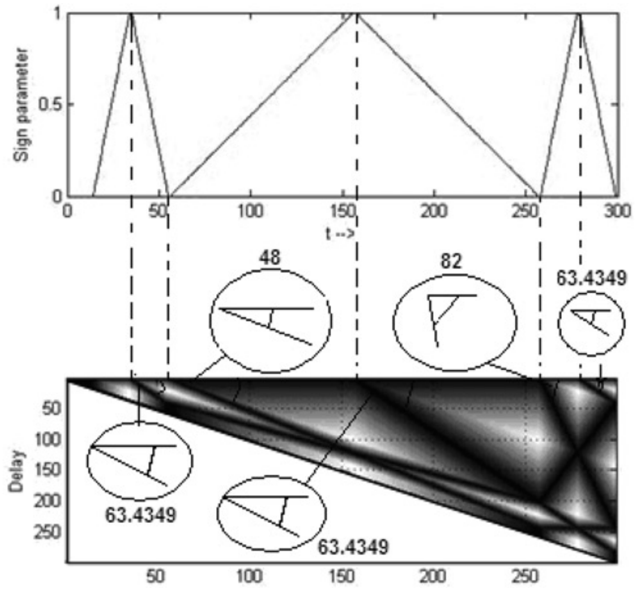


Fig. 21 *DAD signature directional variation segmentation*

the gesture related transitions are extracted for the model matching. As the reported segmentation scheme was proposed in conjunction with a recognition algorithm, no specific detail is found about its segmentation accuracy.

Research [26] presents a direct trajectory segmentation method on 27 SL sentences with minimal velocity and maximum directional angle change. The reported accuracy was 88% with 11.2% false alarm when initial segmentation was subjected to a ‘Naive Bayesian’ classifier.

3.5 *Segmentation signature*

Motion signatures relate the temporal variation of the signer’s body contours with the segmentation of continuous gestures. As shown in Fig. 18, normalised ‘Euclidean distances’ are accumulated from every point belonging to the signer’s body contour to a centroid. Recording of all the distances over a specific period of time generates a patterned surface called ‘motion signature’, shown in Fig. 19 (left). Like any other signature, motion signatures are distinct patterns, which occur at the boundary of two connected gestures (also called compound gestures) [27, 28, 34, 35].

3.5.1 Delayed absolute difference (DAD) signature: DAD signature [36] is similar to the motion signature but instead of using the morphological variations of a signer body, the spatiotemporal variations of the articulator are encoded as distinct segmentation features. DAD signature of a signal quantifies and localises the intra-signal variations which are candidates for the deterministic boundary of a sign-like constancy or pauses, sudden changes and sign reduplication. Fig. 20 presents the segmentation features acquired over a signal with the simulated pauses, directional variations and repetitions. It shows inter-sign pauses, which are transformed into DAD pattern (a black inverted triangle) where the length of its base or height corresponds to the length of the pause segment. Similarly the sign repetitions are short termed so they are simulated as a slightly higher-frequency sinusoid. The resultant repetition patterns in the DAD signatures comprise of black horizontal lines that show the similarity (small differences) of a segment with its previous occurrences.

Another prominent segmentation feature is the abrupt change in the articulator trajectory shown in Fig. 21. DAD patterns for each significant directional variation are encoded as black slanted lines where angle of each line relates to the degree of change in direction.

DAD is a deterministic algorithm for sign boundary detection. Unlike the subjective transcription, which is inherently inconsistent, DAD results are far more robust and reliable. For example, Fig. 22 shows a real signal of three connected gestures of New Zealand SL in a natural sentence. The arrows there show the boundary annotation by an experienced signer. DAD signature extracts some prominent segmentation features at the candidate points (shown in Fig. 23). The length of an inter-sign pause in the every candidate frame can be helpful to retrieve the start of that pause, which can help in retrieving its temporal localisation. Once the ‘start’ and ‘end’ of a pause segment is decided, we can extract the actual gestures from end of previous pause to the start of the next one. By this means the sign components relating only to the linguistically significant units are processed for gesture modelling and recognition. A projection of the detected pauses over the actual signal is given in Fig. 24.

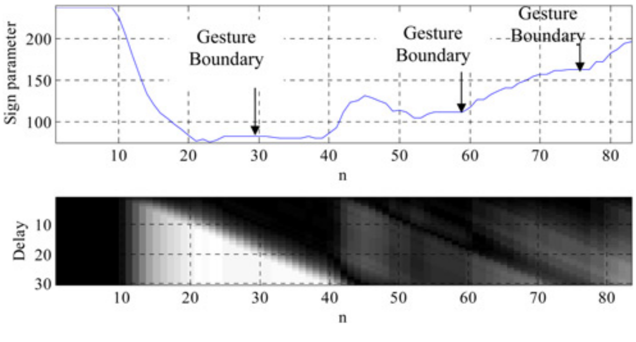


Fig. 22 *Annotated continuous stream*

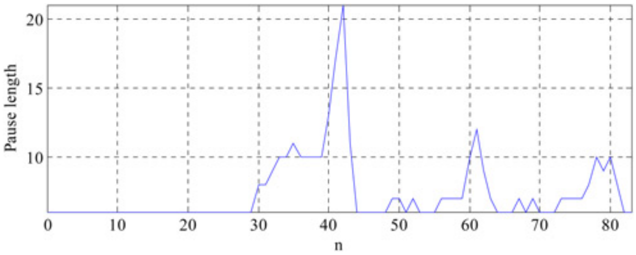


Fig. 23 *Every end point of a pause has a peak that shows its length*

4 Real-time system implementation

In this research, the client/server approach between ‘Windows7’ and ‘LINUX Fedora 16’ was developed for real-time SL recognition. The main objective behind this approach was to utilise positive features of both operating systems. Under the ‘Windows7’, the ‘Visual C++ 2010’ combined with ‘OpenCV 1.1pre’ library can support video processing while providing a powerful graphical user interface. Meanwhile, the Gt^2k for gesture recognition can be

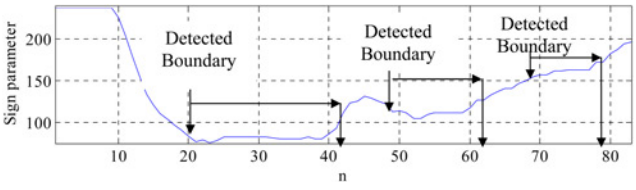
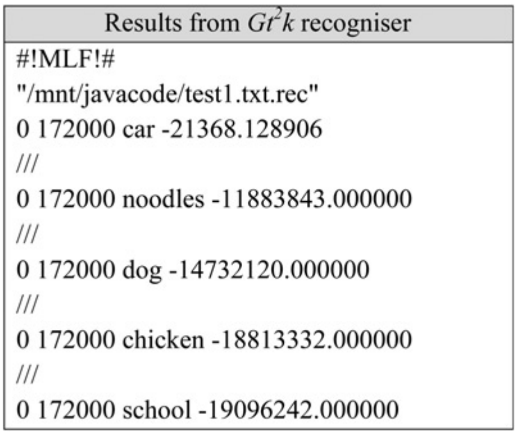


Fig. 24 *Overlaying the detected pause features over the sign parameter (vertical arrows indicate the start and end frames, whereas the horizontal arrows show the length of the quasi-stationary segments)*



Results from Gt^2k recogniser

```
#!MLF!#
"/mnt/javacode/test1.txt.rec"
0 172000 car -21368.128906
///
0 172000 noodles -11883843.000000
///
0 172000 dog -14732120.000000
///
0 172000 chicken -18813332.000000
///
0 172000 school -19096242.000000
```

Fig. 25 *MLF recognition results*

fully supported under ‘LINUX’. Therefore a client/server technology saves significant time during the SL recognition system development.

It also helps to enhance the algorithms in such a way that it is not required to transfer one of the two programs: either the Gt^2k from ‘LINUX Fedora 16’ to ‘Windows7’ or the ‘VC++ 2010’ and ‘OpenCV 1.1pre’ library from ‘Windows7’ to ‘LINUX Fedora 16’. Program ‘VC++ 2010’ cannot be transferred from ‘Windows7’ to ‘LINUX Fedora 16’ operating system because it is a Microsoft product. It offers powerful functionality for developers specifically in developing codes for video and image processing. However, it works only under ‘Windows’ operating system. Another open source product – *Qt* can execute image and video visualising under ‘LINUX’ and could be another alternative for ‘VC++’. Unfortunately as a new programming product *Qt* is offering just limited resources for developers [37].

An alternative method to link the two different operating systems (‘Windows7’ and ‘LINUX Fedora 16’) involves the use of a physical link (cable) between the client and server computers. However, such an approach limits the speed of data transfer and may be affected by potential connection instability. In addition, the cost of this solution is higher since two different computers are required to build such a system.

The reported system was launched on just one powerful computer using two operating systems (‘Windows7’ and ‘LINUX Fedora 16’). This was possible by using the ‘VMware’ workstation software.

4.1 *Sharing files between the client and server*

Client and server are two separate programming modules which communicate through the network while having explicitly distinctive tasks. Usually clients can be categorised into two types: ‘thin client’ and ‘thick client’. A thin client is capable of achieving an acceptable computation performance over wide area networks. However, for the purpose of the reported study, a thick client (‘VC++ 2010’ and ‘OpenCV 1.1pre’ library, which resides in ‘Windows7’) was used. The use of the thick client is significant is this research during the pre-processing of sign videos, which requires high visualisation performance. Similarly, on the server side, Gt^2k is a tool for gesture recognition, which helps human

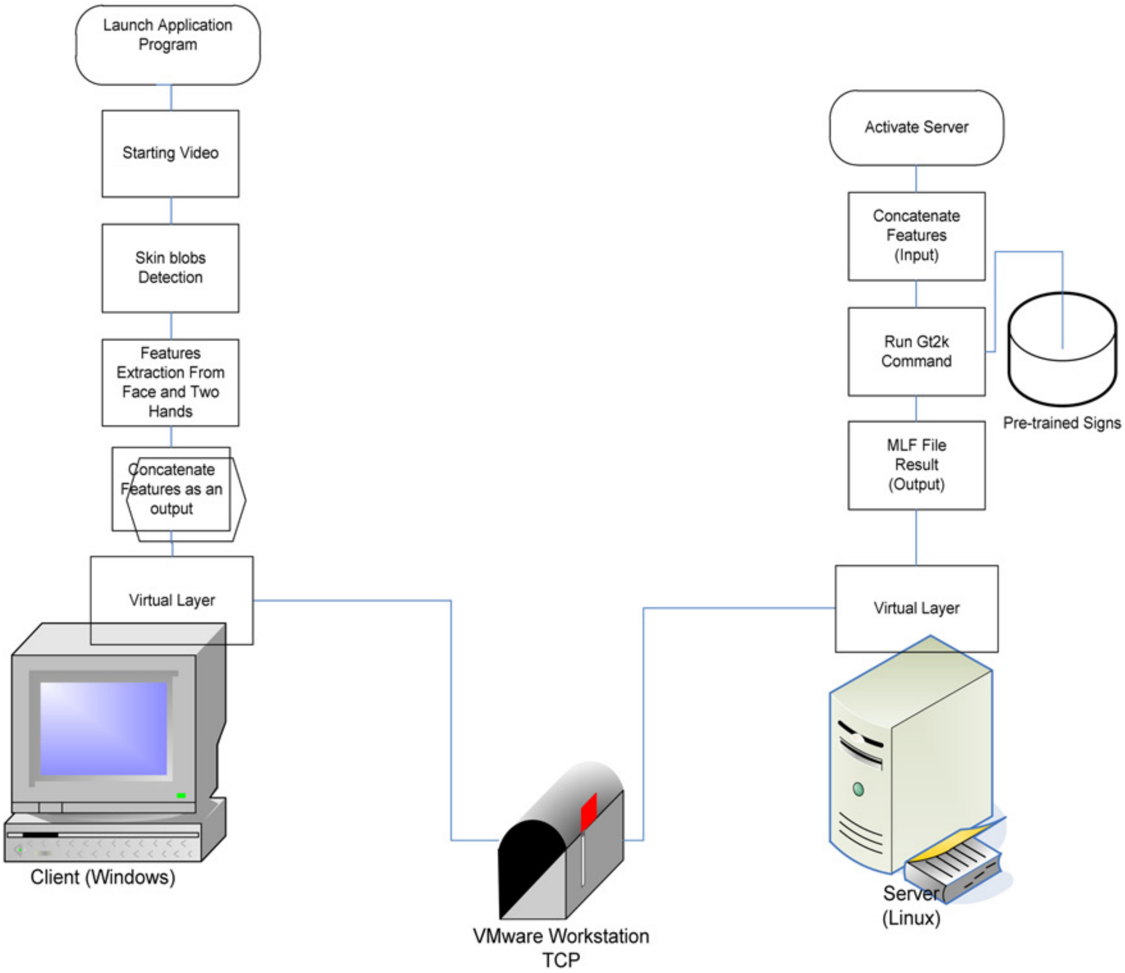


Fig. 26 *Real-time MSL recognition system*

computer interface developers to focus mainly on pre-recognition stages rather than spending time and efforts on building the HMM-based recogniser.

(a) *Thick client:* In this paper the algorithms for face and hands detection, skin blob tracking up to the feature extraction stage were implemented under 'Windows7' environment using 'VC++' 2010 with the help of 'OpenCV 1.1pre' library. To train HMM in an offline mode, feature files were transferred to the Gt^2k for training on a specific number of signs.
(b) *Server for SL recognition:* To achieve the SL recognition using a new non-trained sign, a 'recognize.sh' script can be used from Gt^2k software. Four arguments are to be passed: data files 'signs.txt'; a file to store the recognition results 'results.txt', which will be created automatically; 'options.sh' and the trained HMM model. HMM model is called 'newMacros' and the output is in the form of an MLF. The MLF file contains gestures ranked by their likelihood scores as shown in Fig. 25, which shows MLF for five isolated words. The command for recognition using Gt^2k is

```
Recognize.sh   signs.txt   result.txt   options.sh   NewMacros.
```

As stated earlier, establishing network connection is necessary to begin an interaction between the client and server. It is required to establish communication, which enables the feature collector client to send the features file to the analysing server.

4.2 Thick client and server interface

'VMware' workstation was lunched as a media on 'Windows7' to install 'LINUX Fedora 16'. Following that, a transmission control protocol (TCP) connection type is created between the client and server to transfer the processed video file from the client to the server. Similarly, TCP transfer, the resulted output from the server is inputted to the client as illustrated in Fig. 26.

5 Conclusions

ASLT is an example of a systems leading to a higher life quality. In this research, an automatic Malaysian SL recognition system was developed. The system utilises four stages including: face and hand detection and tracking; HUB detection; feature extraction; and real-time SL recognition. Hands and face are detected by the developed new hybrid method, which combines the appearance-based approach with the skin colour detection algorithm, and which is not restricted by a background. The HUB detection is achieved by applying a new face measurement approach. Various feature vectors from hand shape, hand motion trajectory and hand position are extracted and investigated. During the recognition stage, HMM is used to train and test the developed system using a newly developed Malaysian SL database and implementing new feature matching methods. Meanwhile a real-time SL recognition is implemented using client/server technology between 'Windows7' and 'LINUX Fedora 16'. These interfaces offer additional degrees of mobility and control, which can lead in the future to a betterment and development of portable SL translation devices (which is the ultimate aim of the presented research in the long term).

The overall system processing time for isolated signs into text and/or voice (in English) is <2 s on an upper mid-range commodity personal computer and is shorter on a more powerful machine, thus qualifying for real-time translation. Meanwhile, at this initial prototype implementation stage the average recognition accuracy for 20 isolated signs reached 80% and was at 55% for a total lexicon of 37 words in 20 sentences. It is expected that the accuracy will be significantly increased with the system further development and fine-tuning. Among the additional benefits of the systems is its offering a natural environment for a user where signing can be done by one or both hands, with or without pauses between signs, with flexible velocity of signing, without the burden of wearing any devices and while operating in variable illumination conditions with varying backgrounds.

6 References

[1] MFD, *Malaysian Sign Language*. 2012; Available at http://www.mfd.org.my/public/edu_eSign.asp

[2] MacKenzie I.S.: 'Input devices and interaction techniques for advanced computing, in virtual environments and advanced interface design' (Oxford University Press, Oxford, UK, 1995), pp. 437–470

[3] Pavlovic V.I., Sharma R., Huang T.S.: 'Visual interpretation of hand gestures for human–computer interaction: a review', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1997, **19**, (7), pp. 677–695

[4] Yang M.H., Ahuja N.: 'Recognizing hand gesture using motion trajectories'. IEEE Conf. Computer Vision and Pattern Recognition, IEEE Computer Society, Fort Collins, CO, USA, 1999, vol. 1, pp. 466–483

[5] Grzeszczuk R., Bradski G., Chu M.H., Bouguet J.: 'Stereo based gesture recognition invariant to 3D pose and lighting'. IEEE Conf. Computer Vision and Pattern Recognition, IEEE Computer Society, Head Island, SC, USA, 2000, vol. 1, pp. 826–833

[6] Viola P., Jones M.J.: 'Robust real-time face detection', *Int. J. Comput. Vis.*, 2004, **57**, (2), pp. 137–154

[7] Bilal S., Akmeliawati R., Momoh J., Shafie A.A.: 'Dynamic approach for real-time skin detection', *J. Real-Time Image Process.*, 2012, p. 6

[8] Chang F., Chen C.-J., Lu C.-J.: 'A linear-time component-labeling algorithm using contour tracing technique', *Comput. Vis. Image Underst.*, 2004, **93**, (2), pp. 206–220

[9] Imagawa K., Lu S., Igi S.: 'Color-based hands tracking system for sign language recognition'. Third IEEE Int. Conf. Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 462–467

[10] Bilal S., Akmeliawati R., Salami M.J.E., Shafie A.A., Bouhabba E.M., ET AL.: 'A hybrid method using Haar-like and skin-color algorithm for hand posture detection, recognition and tracking'. Int. Conf. Mechatronics and Automation (ICMA), Xi'an, China, 2010, pp. 934–939

[11] Kilian J.: Simple Image Analysis by Moments, 2001, 8 pp. Available at http://www.scribd.com/doc/39759766/Simple-Image-Analysis-by-Moments

[12] Jusko D.: Full Real Color Wheel Course, 2011. Available at http://www.realcolorwheel.com/human.htm

[13] Bradski G.R.: 'Computer vision face tracking for use in a perceptual user interface', *Intel Technol. J.*, 1998, **2**, (3), pp. 1–15

[14] Bilal S., Akmeliawati R., Shafie A.A., Salami M.J.E.: 'Modelling of human upper body for sign language recognition'. Fifth Int. Conf. Automation, Robotics and Applications (ICARA), Wellington, New Zealand, 2011, pp. 104–108

[15] Starner T.E., Pentland A.: 'Real-time American sign language recognition from video using hidden Markov models'. IEEE Int. Symp. Computer Vision, Coral Gables, FL, USA, 1995, pp. 265–270

[16] Segouat J., Braffort A.: 'Toward modeling sign language coarticulation', in Kopp S., Wachsmuth I. (Eds.): 'Gesture in embodied communication and human–computer interaction' (Springer Berlin Heidelberg, 2010), pp. 325–336

[17] Segouat J.: 'A study of sign language coarticulation', *Spec. Interest Group Accessible Comput. (SIGACCESS)*, 2009, **2009**, (93), pp. 31–38

[18] San-Segundo R., Pardo J.M., Ferreiros J., ET AL.: 'Spoken Spanish generation from sign language', *Interact. Comput.*, 2010, **22**, (2), pp. 123–139

[19] San-Segundo R., Barra R., Cordoba R., ET AL.: 'Speech to sign language translation system for Spanish', *Speech Commun.*, 2008, **50**, (11–12), pp. 1009–1020

[20] Alon J., Athitsos V., Quan Y., Sclaroff S.: 'A unified framework for gesture recognition and spatiotemporal gesture segmentation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009, **21**, pp. 1685–1699

[21] Yang R., Sarkar S., Loeding B.: 'Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2009, **32**, pp. 462–477

[22] Viblis M.K., Kyriakopoulos K.J.: 'Gesture recognition: the gesture segmentation problem', *J. Intell. Robot. Syst.*, 2000, **28**, pp. 151–158

[23] Kahol K., Tripathi P., Panchanathan S., Rikakis T.: 'Gesture segmentation in complex motion sequences'. IEEE Int. Conf. Automatic Face and Gesture Recognition, Seoul, Korea, 2004, vol. 3, pp. II-105–8

[24] Ong S.C.W., Ranganath S.: 'A new probabilistic model for recognizing signs with systematic modulations', Third International Workshop on

Analysis and Modelling of Faces and Gestures, Rio de Janeiro, Brazil, 2007 (*LNCS*, **4778/2007**), pp. 16–30

[25] Ruiduo Y., Sarkar S.: ‘Detecting coarticulation in sign language using conditional random fields’. 18th Int. Conf. Pattern Recognition, 2006, ICPR 2006, 2006, vol. 2, pp. 108–112

[26] Kong W.W., Ranganath S.: ‘Sign language phoneme transcription with rule-based hand trajectory segmentation’, *J. Signal Process. Syst.*, 2010, **59**, (2), pp. 211–222

[27] Li H., Greenspan M.: ‘Segmentation and recognition of continuous gestures’. IEEE Int. Conf. Image Processing, 2007, ICIP 2007, 2007, vol. 1, pp. 365–368

[28] Li H., Greenspan M.: ‘Continuous time-varying gesture segmentation by dynamic time warping of compound gesture models’. Int. Workshop on Human Activity Recognition and Modelling (HARAM2005), 2005, p. 8

[29] Starner T., Pentland A.: ‘Real time American sign language recognition from video using hidden Markov model’. Int. Symp. Computer Vision, Florida, USA, 1995, pp. 265–270

[30] Vogler C.P.: ‘American sign language recognition: reducing the complexity of the task with phoneme-based modeling and parallel hidden Markov models’. PhD dissertation, University of Pennsylvania, USA, p. 172

[31] Guerrero-Curieses A., Rojo-Álvarez J.L., Conde-Pardo P., Landesa-Vazquez I., Ramos-Lopez J., Alba-Castro J.L.: ‘On the performance of kernel methods for skin color segmentation’, *EURASIP J. Adv. Signal Process.*, 2009, **2009**, pp. 1–13

[32] Han J., Awad G., Sutherland A.: ‘Modelling and segmenting subunits for sign language recognition based on hand motion analysis’, *Pattern Recognit. Lett.*, 2009, **30**, (6), pp. 623–633

[33] Liang R.-H., Ming O.: ‘A real-time continuous gesture recognition system for sign language’. IEEE Int. Conf. Automatic Face and Gesture Recognition, Japan, 1998, pp. 558–567

[34] Li H., Greenspan M.: ‘Multi-scale gesture recognition from time-varying contours’. 10th IEEE Int. Conf. Computer Vision, ICCV 2005, 2005, vol. 1, pp. 236–243

[35] Li H., Greenspan M.: ‘Model-based segmentation and recognition of dynamic gestures in continuous video streams’, *Pattern Recognit.*, 2011, **44**, (8), pp. 1614–1628

[36] Khan S., Bailey D.G., Sen Gupta G.: ‘Delayed absolute difference (DAD) signatures of dynamic features for sign language segmentation’. Fifth Int. Conf. Automation, Robotics and Applications (ICARA2011), Wellington, New Zealand, 2011, pp. 109–114

[37] Qt Project. Available at http://www.qt-project.org/