Segregation of meaningful strokes, a pre-requisite for self co-articulation removal in isolated dynamic gestures

Anish Monsley K. | Kuldeep Singh Yadav | Songhita Misra | Taimoor Khan | M. K. Bhuyan | Rabul Hussain Laskar

1 Department of Electronics and Communication Engineering, National Institute of Technology Silchar, Silchar, Assam, India
2 Department of Electronics and Communication Engineering, National Institute of Technology Andhra Pradesh, Tadepalligudem, Andhra Pradesh, India
3 Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, Assam, India

Abstract
Gesture formation, a pre-processing step, has its importance when variations in patterns, scale, and speed come into play. Self co-articulations are intentional movements performed by an individual to complete a gesture, whose presence in the trajectory alters its original meaning. For recognition, most researchers have directly used the trajectory formed along with these self co-articulated strokes, with a few removing it using visible trait-like velocity. Usage of velocity has shortcomings as gesturing in air differs from gesturing over a solid surface; hence, we propose a gesture formation model, which incorporates global and local measures to remove these self co-articulations. The global measure uses Euclidean distance, instantaneous velocity, and polarity calculated from the complete gesture, while the local measure segments the gesture into stroke-level segments by using the minimum–maximum-polarity algorithm and applies the selective bypass rules. The proposed model, when experimented on gestures patterns with premeditated speed variation, has a mean error rate of 0.0069 and 7.40% self co-articulations; individuals’ natural gesticulation has a mean error rate of 0.0371 and 12.07% self co-articulations. Experimentation on each gesture of NITS hand gesture databases showed a relative improvement of 40% (accuracy 97%) over the existing baseline models.

1 INTRODUCTION
Gesture is one of the ways to express a thought by bodily actions. An individual can perform a gesture as additional information to convey what they are trying to say. These gestures help particularly the disabled for whom gesturing may be the primary means of communication. In this regard, the research community has brought forward many developments, where the individual’s speech or gesture movements can perform the intended actions through a human–computer interaction (HCI) machine without unnecessary physical movements. This work is a part of the research that focuses on using hand gesture movements to limit the usage of conventional input peripheral device such as keyboard and mouse. Hand gesture movements can be recorded by non-vision devices such as electrodes directly placed in hands, gloves, inertial sensor or through vision devices such as cameras and Kinect sensor. Though the former provides a more accurate estimation of the intended gesture, the vision-based approach has shown consistent growth in the field of hand gesture (in air) particularly character recognition [1, 2]. The act of gesturing in air includes with it some redundant strokes (pen-up strokes) that need to be removed to get the actual characters. These redundant strokes are intentional movements (ligature motion) made by an individual while gesturing to form the complete meaningful character. Researchers [3] called these redundant intentional movements within a gesture as self co-articulations, while co-articulation is made between two gestures [4, 5]. The number of self co-articulated (SCA) strokes varies with the individual, as a uni-stroke or multi-stroke gesture can be performed even without the usage of SCA strokes. Let us consider a keyboard character ‘A’. This can be gestured with minimum of ‘1–2’ SCA strokes, whereas character ‘B’ can be gestured with ‘0–1’ SCA strokes. Hence, like normal strokes, these SCA strokes are also subjective to the individual
performing it. These strokes do not add/give any further information/meaning to a character, and the presence of SCA strokes may increase the complexity for recognising a gesture, e.g. ‘N’ gesticulated with self co-articulations (see Figure 1). Velocity can be used for segmentation [6, 7] of gesture into strokes and has been used by researchers to remove these SCA strokes [1–3].

It was based on the hypothesis that individuals perform their gesture with self co-articulations at high speed. Gesturing in air varies from gesturing over the solid surface, as an individual performing a gesture (in air) will have SCA strokes with speed equal, greater, or sometimes less than that of normal strokes. Also, people with motion impairment, physically disabled, and elderly cannot perform the gesture with much intentional variation in speed. In these cases, velocity alone cannot be applied as the sole factor for removing these SCA strokes. This motivated us to investigate other distinguishing characteristics present in the gesture and explore new techniques to separate and remove these SCA strokes. The contribution of this paper includes the following:

- Extending the NITS hand gesture database [1] (for virtual keyboard entry) consisting of isolated dynamic gestures to 69 characters with a pattern, speed, and scale variations in the complex (static) background.
- Integration of global and local measures to remove self co-articulation within the gesture.
  - Global (complete strokes)—Time-varying quantities such as displacement and instantaneous velocity are put together with time-independent factors such as polarity using a biased voting pattern to remove SCA strokes.
  - Local (segmented strokes)—SCA stroke segments are bypassed by using a set of formulated selective bypass rules based on the number of stroke segments and the gesture’s initial movements.
- Segmentation of the gesture into stroke-level segments is done by the proposed minimum-maximum-polarity (mMPo) algorithm for applying the selective bypass rules.

The remainder of this paper is organised as follows. Section 2 briefs related works. Section 3 describes the developed database. Section 4 details about the proposed system model. Section 5 elaborates the results and discussion. Section 6 summarises the conclusion and highlights the future direction of work.

2 RELATED WORKS

Gesture recognition is a broad term including, in itself, the following—gesture object identification, gesture formation, and gesture recognition. Researchers working in this area of gesture recognition focus either on the gesture object identification or on gesture recognition, as the input data is a static gesture (only spatial information) where an individual poses a gesture (identification + formation) and recognises it using conventional feature-based learning models or the deep architecture models. But in the case of dynamic gestures, gesturing is a time series of static frames (spatio-temporal information), and in it, gesture (trajectory) formation is also an important step before recognition. The works of literature reviewed below are analysed in aspect to object detection, data acquisition technique, gesture formation and input data used in their respective work.

For gesture recognition, two researchers developed the NITS hand gesture database consisting of 40 characters, which were collected from 20 individuals. They used a red marker as a gesture object and a 2D camera for capturing the gestures. They calculated velocity between each frame and used the average velocity as a threshold to separate the normal strokes from SCA strokes. An improvement of \( \sim 20\% \) accuracy was attained when SCA strokes were detected and used as a feature to recognise each gesture [3]. Researchers detected bare hand in 2D HSV and 2D YCbCr colour space to form the gesture trajectory for isolated and continuous gestures made using numerals from 0 to 9 [8]. Twenty-two individuals were asked to gesture the 26 uppercase English alphabets and 10 words in a single continuous stroke. Based on the gesture motions, hidden Markov models (HMMs) for each character and HMMs with gesture motion for word level were developed to recognise them [9]. Utilising a sliding window with a step size of 10 frames, rules for detecting the gesturing in air events—writing and non-writing—were developed. For detection, seven features were extracted from both the events, and the Gaussian mixture model was used to classify. For recognition, 2D position and velocity were extracted from the detected frames, normalised, and then used as features for the HMM model [10]. To overcome the influence of speed variations in the gestures, researchers tried to alter the threshold velocity by including 50\% of average velocity to it. Furthermore, to limit the distortions caused by the slight hand movements, frames with velocity lesser than 50\% of the average velocity were also removed. These two steps helped mitigate the presence of SCA strokes to some extent [11, 12]. To make use of movements in the hand muscle for gesture recognition, researchers measured the muscle activity using electrodes during hand gesturing to form instantaneous high-density surface electromyography image. Databases such as CSL-HDEM G, NinaPro, and CapgMyo was used in their work. Statistics were calculated from the high-density scanning electron microscopic image to train the machine learning models for classification [13]. After designing a wireless inertial motion unit sensor, researchers [14] tried to capture the 3D handwritten gesture performed in a solid surface. Twenty volunteers were asked to gesture English alphabets (lowercase) and numbers from 0 to 9 in one direction as a single stroke.
Words were separated from non-words using the device’s inbuilt motions metrics such as acceleration, angular velocity, and 3D attitude (orientation).

In [15], researchers normalised the speed variation using the dynamic time warping (DTW) technique. Then, they used Euclidean distance to match the trajectory points to a template, which was a trajectory formed from averaging every gesture trajectory points. Researchers in [16] detected fingertip through the 3D Kinect sensor and gave instructions as a static gesture to a virtual text keyboard with the assistance of an inbuilt flick keypad. In [17], 3D text gestured in the air with a single stroke was captured using a leap motion sensor. Then, researchers segmented these words using a sliding window with a heuristic search for finding the end of each word. In total, 320 sentences recorded from eight users were used for their work. Instead of forming a complete trajectory of the gesture, they [18] detected hand from each trajectory points of a frame with the help of principal component analysis, which reduced the dimensions. They used the MSR hand gestures dataset and the VIVA challenge dataset. Researchers [19] proposed the distance-weighted-curvature-entropy algorithm to detect and track the fingertip of a hand to form a complete trajectory of the gesture. 1800 gesture samples of numbers from 0 to 9 and 26 lowercase English alphabets were used for their work. Utilising the reflectance characteristics of a hand from an IR image emitted from a leap motion device, researchers [20] captured 16 static and four dynamic gestures. They used a combination of two-hand pose for representing dynamic gestures. Researchers [21] detected fingertip and captured the gesture and its trajectory using the Intel RealSense SR300 camera. The trajectory was normalised using nearest neighbour point and root point algorithms. RTD and 6MDG trajectory datasets were used for their work. In [22], they extended the work of [16] to detect hand in 2D HSV model space from a simple background. They had introduced seven new hand static gestures to control the virtual keyboard and to feed the input. To capture the motion data, researchers [23] made individuals to wear a wristband while gesticulating 26 English alphabets (Uppercase). The gesticulation act was based on the user-dependent model and the user-independent model. Accelerometer and gyroscope features captured from the wrist band were used to train the DTW + k-nearest neighbours and convolutional neural networks for classification. Researchers [24] proposed a long short-term memory (LSTM)-based dynamic probability method for segmentation and recognition of static continuous gestures. The LSTM model was trained with invalid gestures (short and long trajectories) and valid gestures (Arabic numbers). Segmentation of the invalid sub-gestures was performed by merging only the input trajectory points with a trained model. The merger operation was based on the highest probability, i.e. the largest vector point of LSTM was taken as a confidence measure. Based on the number of SCA strokes in a gesture, researchers [25] developed a hierarchical model to distinguish 58 gestures (26 alphabets, 0–9 numbers, and 22 symbols). Then, trajectory-based and image-based features were extracted from the model and were used to train the Voronoi-diagram-based classifier and neuro-fuzzy for recognition.

To sum up, researchers have used the gesturing objects like a fingertip, data gloves, electrodes, colour marker, hand, inertial motion sensors such as wireless-inertial-measurement unit, palm’s GEOS-based devices, and wrist bands. Input acquisition devices such as a 3D-Kinect sensor, leap motion sensor, time-of-flight camera, infrared camera, stereo camera, and 2D camera were used to capture the gesturing act. Once gesture object is detected, they have either used the trajectory formed directly or matched it with a predefined template/character-set/graffiti/uni-stroke symbols by calculating the DTW distance or finding the absolute error/differential programming distance, the minimum of which is considered to form the gesture [26, 27]. This indicates that in both the cases, the pattern of the gesture is predefined/unitary in a single direction with fewer self co-articulations. Hence, these techniques would fall short when the gesticulating pattern for a gesture varies, involving directional changes and more self co-articulations to complete the gesture. Thus, the above techniques prevent the individual from using her/his inherent gesticulation style to gesticulate in the air for HCI. On the other hand, few researchers used velocity to remove these self co-articulations with a notion that these are of much higher speed. But when individuals gesture over the air, there is negligible speed variations between the normal and the SCA strokes unlike gesturing over the solid surface (pen-down and pen-up strokes) where there exist visible speed variations. Hence, when an individual gesticulates in the air with her/his innate gesticulation speed, velocity cannot be used to remove these self co-articulations.

In all of these cases, researchers have used a minimal number of characters to demonstrate their work. Hence, there is a need to: a) develop a database with more number of characters which are gestured in a natural way incorporating the variations in pattern, scale, and speed; b) find an alternative solution to remove this self co-articulation from the conventional velocity-dependent and template matching techniques; and c) make the system model robust to variations in pattern and scale. In this work, we chose to use appearance-based gesture object, i.e. red colour marker, and a 2D camera for capturing the gesture in a complex environment and extend the NITS hand gesture database (for virtual keyboard entry).

3 DATABASE DEVELOPMENT

In the NITS hand gesture database (for virtual keyboard entry), the volunteers with consent are asked to perform 95 keyboard characters, i.e. the isolated dynamic gestures. Each volunteer is instructed to perform the 95 gestures using a marker whose rear end is red coloured (size of 1 cm). Gesturing in the air should be within the purview of a 2D camera (Logitech 930e webcam with a frame rate (FR) of 30 fps) in a natural way they usually write. Thus, the gesturing act has no restraint on: i) gesture pattern; ii) speed of gesticulation; and iii) scale of gesture. The volunteers were asked to perform again if their gesture was not recorded properly. Each gesture was saved as a separate video with a filename, e.g. ‘Re0001M001N00001’. In this nomenclature, position ‘1-2’ indicates red colour marker,
position ‘3-6’ indicates volunteer’s identity number, position ‘7’ refers to the gender, position ‘8-10’ indicates the gesture number, position ‘11’ A/N/S indicates what is that gesture, i.e., alphabets/numbers/symbols, and position ‘12-16’ indicates the video number.

Details of the 95 gestures are (26 English alphabets in capital, 26 English alphabets in smaller, 10 Numbers (0–9), 33 Symbols (addition ‘+’, subtraction ‘−’, multiplication ‘∗’, division ‘÷’, tildes ‘~’, at sign ‘@’, equal to ‘=’, caret ‘^’, ampersand ‘&’, left curly bracket ‘{’, right curly bracket ‘}’, left square bracket ‘[’, and right square bracket ‘]’, left round bracket ‘(’, right round bracket ‘)’, backslash ‘\’, question mark ‘?’, percentage ‘%’, lesser than ‘<’, greater than ‘>’, dollar sign ‘$’, exclamation sign ‘!’, slash ‘/’, dot ‘.’, comma ‘,’ colon ‘:’, semi-colon ‘;’, hash ‘#’, left arrow ‘←’, right arrow ‘→’, top arrow ‘↑’, bottom arrow ‘↓’, and backspace ‘⌫’). The database at present has 5320 gestures videos collected from 56 volunteers—NITS (45 males and 10 females), of which 3 are left-handed and ‘1’ male is from general public. In this research work, we will use 69 characters leaving out the lowercase English alphabets.

4 PROPOSED METHOD

In this research work, all the gesture videos are pre-processed using open-source video editing software. This is performed to make the red coloured marker, which is detected in the first frame be the starting coordinate point and the last frame be the last coordinate point for the gesture.

4.1 Red marker detection

The first step of gesture identification is to detect the red coloured marker, for which we have followed an approach similar to [12]. All the red coloured objects are first detected from each frame by subtracting the green (Ig) and the blue channel (Ib) from the red channel (Ir) using (1). Then, to obtain the luminance of the red coloured objects, the red channel is subtracted from its grey channel (GR) by (2). Finally, using (3), the Otsu threshold minimises within-class and maximises between-class variance and groups them [28], which would result in obtaining our desired region of interest (ROI) (red coloured marker). This ROI’s centroid coordinate is detected in each frame using (4), which would be tracked to form the complete gesture [see Figure 2(a)]. Then, each coordinate point is checked with its neighbouring coordinates; if the difference of them exceeds a limit, we replace these coordinates with its immediate succedent coordinate subtracted from a random integer through (5) [see Figure 2(b)]. This will make the gesture to be in its intended path. By experimentation, the limit of 50 was chosen, as it worked for all the different scales used for gesturing. The trajectory is further smoothed (SM) by averaging with its immediate neighbours by (6) [see Figure 2(c)]. Figure 2 illustrates the working of (4)–(6) for a gesture ‘F’.

\[
R(i, j) = I_r(i, j) - (I_g(i, j) + I_b(i, j)) \quad (1)
\]

\[
L(i, j) = R(i, j) - G_R(i, j) \quad (2)
\]

\[
G_R(i, j) = 0.299 * R(i, j) \quad (3)
\]

\[
CE_x = \sum_{p=a}^{b} \frac{j(p)}{b} ; \quad CE_y = \sum_{q=a}^{b} \frac{i(q)}{b} \quad (4)
\]

where \(i = 1\) to \(N\); \(j = 1\) to \(M\).

\[
RM(k, l) = \begin{cases} \left[ CE_x(Fr) ; CE_y(Fr) \right] & \text{if } CE_x(Fr + 1) - CE_x(Fr) < 50 \\ CE_y(Fr + 1) - CE_y(Fr) < 50 \end{cases} \quad (5)
\]

\[
SM(X, Y) = \frac{RM(k - 1) + RM(k) + RM(k + 1)}{3} \quad (6)
\]

where \(k = 2\) to \(Fr - 1\); \(l = 2\) to \(Fr - 1\); \(X = 1\) to \(Fr - 2\); \(Y = 1\) to \(Fr - 2\).

\[\]

\[\]

\[\]

\[\]

\[\]

\[\]

\[\]

\[\]

\[\]

\[\]

\[\]

\[\]

\[\]

\[\]
4.2 Removal of SCA stroke

Self co-articulations in gesture are a part of intentional movements that are needed to complete the gesture. The red marker detection in each frame will include these self co-articulations, which can distort the actual intended meaning of the character (see Figures 1 and 2). Statistical analysis on the trajectory points revealed that the number of coordinate points and the distance between successive coordinate points vary in the normal and SCA stroke; instantaneous velocity (metre/second) can be formulated using pixel density (pixel per inch—PPI), scale information, and distinct polarity observance from the gradient of strokes. Visual observation of gesticulating act showed that, if the gesture stroke does not begin at its conventional ground truth (GT) position, there is a natural tendency to make use of self co-articulations every second time to complete the gesture. However, to make use of this proposition, the trajectory must be segmented into strokes segments, for which an mMPo algorithm is proposed. Based on the number of segmented strokes segments and initial gesture movements (horizontal/vertical) information to choose selective bypass rules (hard threshold) have been formed. Integration of two approaches is being utilised for the removal of SCA strokes, and the proposed system model is shown in Figure 3.

**Euclidean distance (ED):** Displacement is the change in position of vectors; this is one of the primitive visible characteristics that can be obtained from the trajectory points. To measure the distance of each successive 2D coordinate points in the trajectory, using (7), the Euclidean distance is calculated. This is one of the commonly used distance measures in gesture recognition points (pixels/frame), the frame rate of video (frame/second), and scale (metre/pixel). Pixel density (PD) was used to calculate the scale information

$$ID_{sd} (no) = \sqrt{(SM(X + 1) - SM(X))^2 + (SM(Y + 1) - SM(Y))^2} \quad (8)$$

$$ID_{pd} (no) = \sqrt{(SM(Y + 1) - SM(Y))^2} \quad (9)$$

$$FR = Number of frames/second \quad (10)$$

$$PD_{inches} = \frac{W^2 + H^2}{D} \quad (11)$$

$$PD_{metre} = PD_{inches} * 39.3701 \quad (12)$$

$$Scale = 1/PD_{metre} \quad (13)$$

$$IV_{sd} (no) = ID_{sd} (no) * FR * Scale \quad (14)$$

$$IV_{pd} (no) = ID_{pd} (no) * FR * Scale \quad (15)$$

$$TV_{xy} (no) = ED(no) * FR * Scale \quad (16)$$

where no = 1 to Fr − 2; X = 1 to Fr − 2; Y = 1 to Fr − 2.

The subscripts sd, pd, and xy represent the coordinate points of x-axis, y-axis, and both x-axis and y-axis. The FR represents the number of frames per second in the input video (FR = ~30 fps). Here, W, H, and D represent the width, height, and diagonal length of the display unit. The constant 39.3701 corresponds to metric of PPI to pixel per metre conversion.

**Polarity (Po):** Analysis of the gradient of each consecutive 2D coordinate points revealed that some of the normal strokes and the SCA strokes are of opposite polarity. We used the gesture initial movements (horizontal/vertical) information to choose the polarity that corresponds to the respective strokes by using (19). This concept is similar to the zero-crossing rate of speech signal processing [29]

$$P_{sd} (no) = SM(X + 1) - SM(X) \quad (17)$$

$$P_{pd} (no) = SM(Y + 1) - SM(Y) \quad (18)$$

$$P_{xy} (no) = \begin{cases} SM(Y(1)) - SM(Y(M)) < 0 \\ SM(X(1)) - SM(X(M)) > 0 \end{cases} \quad (19)$$

where X = 1 : Fr − 2; Y = 1 to Fr − 2; no = 1 to Fr − 2.
The variable \( P \) is the output from the gradient operation along the x-axis \( (P_{dx}) \) and along the y-axis \( (P_{dy}) \). \( P \) would consist of coordinate points from the normal stroke segments. The value of \( M \) is chosen to be 10/15 depending upon the scale of the gesture. To use polarity for removing self-co-articulation, the original starting and ending coordinate point of the gesture is again kept as start and end coordinate point along with \( P \) coordinates. In some cases, removal of the SCA stroke resulted in either removing the first stroke or last stroke, as they were in the same polarity. The above-mentioned condition will help us overcome this limitation of polarity at the time of gesture formation. Thus, the resultant formed characteristic ‘Po’ is invariant to the pattern, as for each pattern the consecutive polarity would also change depending on the initial gesture movements. As it depends on the spatial coordinates only, dependence on velocity can be minimised. In the case of scale variation, the polarity would remain the same irrespective of the size. Thus, polarity will support other visible characteristics like velocity and displacement for the detection and removal of self co-articulations.

**Biased voting** Time-dependent characteristics such as Euclidean distance and instantaneous velocity and time-independent characteristics such as polarity are allotted with equal weights of 0.5. A biased vote is cast to make the system model less dependent on speed variation of the gesticulating pattern. A coordinate point to get selected needs a threshold greater 0.5 suggesting that all candidates must necessarily satisfy the polarity criteria. It was found experimentally that this voting pattern will facilitate in getting the correct coordinates when the strokes of the gestures are performed irrespective of speed variations. Figure 4 shows an example of biased voting’s candidate selection.

**Local approach:** The second model explores the local characteristics, wherein a gesture will be segmented into individual stroke segments. To understand the motor control complexity of an individual while gesturing an unfamiliar character, researchers developed a model, which decomposes the gesture strokes into curves, lines, and corners (CLC). This CLC model used the radius of curvature and various line fitting models for the gesture segmentation. Then, they calculated the production and execution time of each stroke, which exhibited the complexity rate [30]. Using the Niclcon database, researchers calculated 758 existing stroke-level features from pen-down strokes and pen-up strokes. Using classifiers such as support vector machine (SVM), multi-layered perceptron, and template matching DTW technique, the performance of the selected feature set was analysed, and gestures were recognised [31]. For a multi-stroke gesture, researchers generated possible stroke orders and the direction of each uni-stroke using permutation \( (2^N) \) combination for \( N \)-strokes. The most likely uni-stroke was selected through speed optimisation, i.e. angle of the starting point. Then, each of the uni-stroke was joined using the endpoints of successive strokes and matched with the multi-stroke template for recognition [32]. Based on the maximum curvature, minimum velocity, and x-axis coordinate endpoint, researchers segmented a string of numbers into each stroke segment. Then, from each stroke segment, geodesic distance and nine binary geometric features were extracted. A class score is allotted to each segment using the SVM to discriminate the strokes. Dynamic programming was then used to choose the optimal path, i.e. segments with maximum score [33].

To overcome the influence of gravitation as a noise, researchers calculated the standard deviation of tri-axial acceleration signal and used it as a threshold to segment continuous uni-stroke gestures. By analysing the standard deviation of each individual’s gestures, a user-dependent threshold was developed by their online system model [34]. Researchers customised a fully connected network (FCN) with dynamic weighted binary cross-entropy loss function to classify the splitting points (start and end of character) and non-splitting points (end of preceding and beginning of next character) of a gesture. The output of the FCN was then binarised, and the adjacent middle splitting points were formed as pairs (start and end of character), while the other points were removed [35]. Using the gesture phase segmentation dataset, they classified the gesturing phase (act) into segments (rest position, preparation, pre-stroke, strokes, post-stroke, retraction, and rest position) based on spatial and temporal features obtained from each frame using multi-layer perceptron [36]. The above literature consists of the works done on the segmentation of hand gestures; it shows that researchers have focused more on segmenting/to separate different gestures for the further stages of gesture recognition. The amount of research done on segmentation within a gesture, i.e. stroke level is at a bare minimum, primarily because of the steady progress in the machine learning. However, stroke-level segmentation is still an area of research yet to be focused, as it can help in faster recognition, i.e. instant character recognition by the use of stroke segments. This motivated us to develop an algorithm to segment an isolated dynamic gesture’s trajectory into individual stroke segments. This segmentation algorithm works by finding the line and curve segments in stroke level by an iterative search.

**mMPo segmentation:** The gesture is segmented into individual stroke segments by combining min–max-polarity characteristics of the gesture. The algorithm works by first finding the minimum and maximum coordinate point (y-axis) of the trajectory formed as an initial guess. As mentioned earlier, based on the gesturing act, a gesture can be made with/without self co-articulations. It was observed from the collected data that the possibility of using the SCA stroke to gesture the first stroke is more. Hence, detecting the first stroke, i.e. start to end coordinate of the first stroke, is important. This made us introduce
the min–max-polarity rule for detecting the endpoints of each stroke. To elaborate more, if self co-articulations was used to make the first stroke, the minimum point will be selected as the endpoint coordinate for the first stroke. Then, the maximum point will be used as the endpoint of the second stroke. Conversely, if self co-articulation was not used for gesturing the first stroke, the algorithm will choose the maximum point as the end point of the first stroke and minimum point as the endpoint of the second stroke. Once the two strokes are found out, the remaining strokes will be segmented by taking the gradient of each successive trajectory points. The first instance that may be either along x-axis or y-axis, where a polarity change occurs in the gradients will be endpoint of the third stroke; similarly, the process is iteratively done to find the gradient of remaining coordinate points and checked for each first instant polarity change, which will be the endpoints of each successive strokes. The algorithm also checks each time when an endpoint is detected, whether it matches the endpoint of the trajectory, and if so, it ends. It is also to be noted that polarity change within the detected minimum and maximum will also be checked, and if found so, the same process mentioned above will be applied. The flowchart of the proposed algorithm developed is shown in Figure 5. The working of the proposed algorithm is illustrated pictorially using gesture ‘A’ in Figure 6. Once the strokes are segmented, they are fed to a selective bypass to get the normal strokes.

**Selective bypass rules:** Usage of SCA strokes to complete a gesture is subjective; however, it can be observed from the data collected as well as from NITS hand gesture—VI [1] that the number of patterns is finite and gesticulating patterns can be grouped based on the number of strokes and stroke segments. For example, letter ‘E’ has various gesture patterns, but the number of stroke segments resulting from a set of patterns is same, giving insights that the same gesture with different patterns and a different gesture can be grouped based on the number of stroke segments. On further classifying these gestures based on their initial movements revealed that the location of SCA stroke was also identical. Based on the above preliminary observation about the initial movements of the gesture, the position of SCA stroke, and the number of stroke segments obtained from the proposed algorithm, we formulated seven hard threshold rules, which will help bypass those stroke segments that are SCA (see Table 1). Also, the rationale behind the rule-based selection of segments is that utilisation of crafted features and machine learning model for a pre-processing stage like self co-articulation removal will help avoid computations that can benefit us in later stages. The rules we designed are shown in Table 1.

5 | RESULT AND DISCUSSION

For the performance analysis, the minimum number of stroke segments required to form a gesture is taken as GT, and the number of normal strokes segments in the gesture formed by the proposed model gives the arrived truth (AT). The difference in GT and AT would give the missed normal segments (error), and that calculated for all the gesture patterns would give its mean error

\[
\text{Mean error rate (MER)} = \frac{1}{\text{Number of ground truth segments}} \sum_{\text{Number of gestures}} (\text{GT} - \text{AT}).
\]
TABLE 1  Rules for bypassing the SCA strokes

| Rule | Number of segments | Gestures with various patterns in each group | Initial movement is towards | Bypass segments |
|------|-------------------|---------------------------------------------|---------------------------|-----------------|
| 1    | ≤2                | 38                                          | North                     | 0               |
|      |                   |                                             | South                     | 0               |
|      |                   |                                             | East                      | 0               |
|      |                   |                                             | West                      | 0               |
| 2    | 3                 | 29                                          | North                     | 2               |
|      |                   |                                             | South                     | 2               |
|      |                   |                                             | East                      | 2               |
|      |                   |                                             | West                      | 0               |
| 3    | 4                 | 29                                          | North                     | 3               |
|      |                   |                                             | South                     | 2               |
|      |                   |                                             | East                      | 2               |
|      |                   |                                             | West                      | 3               |
| 4    | 5                 | 21                                          | North                     | 2,4             |
|      |                   |                                             | South                     | 2,4             |
|      |                   |                                             | East                      | 2,4             |
|      |                   |                                             | West                      | 4               |
| 5    | 6                 | 7                                           | North                     | 3,5             |
|      |                   |                                             | South                     | 3,5             |
|      |                   |                                             | East                      | 2,5             |
|      |                   |                                             | West                      | 4               |
| 6    | 7                 | 5                                           | North                     | 0               |
|      |                   |                                             | South                     | 2,4,6           |
|      |                   |                                             | East                      | 2,4,6           |
|      |                   |                                             | West                      | 0               |

The MER does not convey the presence of SCA strokes as only the normal stroke segments from the gesture formed are considered. To analyse the performance of the proposed model in terms of overall self co-articulation removed, a ratio of the number of SCA stroke segments present in the gestures formed to the total number of SCA strokes segments in the gesture patterns is taken as another measure.

5.1  Performance of the mMPo algorithm

To check the segmentation performance, the MER was calculated for each pattern of the gesture under the extended version of the NITS database. In this case, GT would be the stroke segment along with the observed SCA strokes for different patterns, and AT would be the resultant stroke segments from the ‘mMPo’ algorithm. Based on the number of patterns in each of the 69 gestures, an average ratio of 1.8987 patterns per gesture with the range of 1–7 patterns was observed. The ‘mMPo’ algorithm for segmenting isolated gesture had an MER of 0.0357 with few individuals’ gesture exhibiting under segmentation when the gesture scale was reduced to (1/3)rd of the normal scale. This is because of the threshold we kept, i.e. the first polarity change should come either after the next 10th frame in the y-axis coordinate and the next 10th frame in the x-axis coordinate. To elaborate further, as we use polarity to segment strokes, even as slight movement (along the x-axis or the y-axis) in the gesture will cause a change in polarity, which will result in over-segmenting even a line. As we wanted the gesture formed to be in its natural state as much as possible, we did not go for any stringent smoothening process, which may be one of the reasons for the error rate. To further clarify it, when we reduced the threshold of the x-axis coordinate further to 5, the segmentation worked properly for that scale. In some cases, gestures ‘percentage’ and ‘division’ show undersegmentation as the circle used in the gesture was not as distinguishable from a dot, and in ‘question mark’ and ‘exclamation mark’, the self co-articulations were aligned with the straight-line segment of the gesture, which made it difficult for segmentation. Figure 7 shows some of the segmentation output from the ‘mMPo’ algorithm. For better visualisation, the segmented outputs are put together, and each segment is shown in different colour and texture.

Two experiments were conducted on the proposed model of gesture formation, and the resultant gesture was compared with the GT in the aspect of normal and SCA stroke segment’s presence.

5.2  Experiment 1: With a reference pattern

From the NITS hand gesture database—VI [1] consisting of 58 gestures, 2913 videos were selected, where individuals were instructed to perform a gesture with intentional speed variation between the normal and SCA strokes. This experiment is to test the proposed model’s performance when each gesture has: i) SCA strokes with a velocity higher than normal strokes; ii) a definite scale; and iii) a fixed pattern. Figure 8 shows the velocity along the x-axis and the y-axis when gesture ‘A’ is performed with reference.

As these gestures had intentional speed variation, SCA stroke will exhibit maximal displacement and the Euclidean distance as a measure exhibited this displacement. Gestures with stroke segments having displacement were detected by (7), and the maximum of which was removed. The same was the case with
instantaneous velocity; using (16), stroke segments with maximal velocity were removed. Once we identified the coordinate point that showed the above factor, its immediate four neighbours that corresponded to the SCA stroke were also removed. It was found experimentally that the four immediate neighbours contributed to self co-articulations in most cases. For polarity, the gradient of all the gesture coordinates was calculated using (17) and (18), and by (19), gestures with normal strokes were obtained. Then, to check the performance of global measure, using (20), the MER was calculated for ‘ED’, ‘IV’, and ‘Po’. An MER of 0.0209, 0.0279, and 0.1818, respectively, was achieved when AT segments were matched with GT segments. As the chosen input had intentional speed changes, ‘ED’ and ‘IV’ removed the SCA strokes to a certain extent compared to ‘Po’. From now onwards, the number succeeding the underscore next to a gesture would represent the total number of segments in that gesture pattern. This will help in visualising the ‘OR’ operation performed at a later stage and to understand the pattern variations in the gesture (see the Appendix). In gestures ‘I_5’, ‘J_4’, and ‘T_3’, Euclidean distance and instantaneous velocity removed a part of the normal segment along with self co-articulations. This is because, an individual’s gesticulation speed is subjective, which makes the number of coordinate points used in SCA strokes vary from 4 to 10. The number of coordinate point in the self co-articulations for these gestures is less than four, which resulted in the removal of a part of the normal segment.

Gestures fed to the polarity block resulted in ‘G_4’, ‘M_4’, ‘N_3’, ‘O_2’, ‘Q_3’, ‘R_3’, ‘W_4’, ‘Z_3’, ‘zero_2’, ‘2_3’, ‘4_3’, ‘5_3’, and ‘question mark_4’ losing one of its normal segment’s coordinate points, whereas ‘8_4’, ‘@_4’, and ‘&_4’ lost two segment’s coordinate points, and ‘percentage_7’ and ‘division_7’ lost three segment’s coordinate points. This is because ‘Po’ by (19) detected a change in the polarity of those stroke segments and removed them. Gestures ‘question mark_4’ and ‘exclamation mark_3’ still had one SCA stroke, as ‘Po’ was not able to distinguish the SCA and normal stroke. The presence of SCA strokes indicates that if SCA strokes are with the same polarity of normal stroke, ‘Po’ fails to remove them. When the coordinate points of the three blocks are fed to biased vote, those coordinates that satisfied the threshold greater than 0.5 were passed through. Then, the MER was calculated from the segments of the biased vote with GT. Thus, by global measure, an MER of 0.1958 was obtained, and the formed gestures had 3.70% self co-articulations. The increase in the MER was due to the loss of normal segments in global measure. When the gesture patterns of experiment 1 were fed to local measure, the ‘mMPo’ algorithm segmented them and fed to the selective bypass rules, which bypassed the SCA strokes. The total number of patterns observed in this reference dataset was 58, of which the rules correctly bypassed SCA strokes in 75.86% of gesture patterns.

In gestures ‘G_4’, ‘M_4’, ‘N_3’, ‘R_3’, ‘W_4’, ‘Z_3’, ‘2_3’, ‘8_4’, ‘@_4’, ‘&_4’, ‘percentage_7’, ‘division_7’, ‘question mark_4’, and ‘exclamation mark_3’, the rules failed by bypassing the wrong segments. An example of a failure case is gesture ‘R_3’, where rule 2—North bypasses the second segment (see Table 1), which results in gesture ‘R_3’ losing a normal segment. The gestures formed with the remaining segments were matched with the GT and the MER was found to be 0.1258 and had 7.40% of self co-articulations in the gestures formed. The increase in the MER was due to loss of normal stroke segments in gesture with no self co-articulations, and this indicates that the performance of local measure worsens for gestures with no self co-articulations. The selected coordinates from both the global and local measure are fed to a union ‘OR’, which performs a union operation to select all of them. Gesture ‘G_4’ had lost its third segment coordinate points from local measure when fed to the union ‘OR’ gets the complete gesture with all the four-stroke segments. As ‘G_4’ had all the segment coordinates from the global measure, the union operator provided the coordinates, which were removed in local measure. Similar was the case for the gestures ‘M_4’, ‘N_3’, ‘O_2’, ‘Q_3’, ‘R_3’, ‘W_4’, ‘Z_3’, ‘zero_2’, ‘2_3’, ‘4_3’, ‘5_3’, ‘8_4’, ‘@_4’, ‘&_4’, ‘percentage_7’, ‘division_7’, ‘question mark_4’, and ‘exclamation mark_3’ (see the Appendix). Also, gestures ‘I_5’, ‘J_4’, ‘T_3’, which had lost a part of its normal segment (centre segment) from the global measure, were able to form the complete gesture with the coordinate points of the same gesture from the local measure. Gestures ‘percentage_7’ and ‘division_7’ which had lost its second, fourth, and sixth segments coordinate points from the local measure and second, fifth segment coordinate points from the global measure was able to form the complete gesture. Gestures ‘question mark_4’ and ‘exclamation mark_3’, which had lost its second segment from the local measure, now formed the complete gesture. But one SCA stroke segment was still present in the gestures ‘percentage_7’ and ‘division_7’, ‘question mark_4’, and ‘exclamation mark_3’. The proposed system model at the end of experiment 1 had an MER of 0.0069 and 7.40% of self co-articulations. The rationale behind the selection of all the coordinates, i.e. the union ‘OR’ operation, is that the proposed system model should remove the self co-articulations even at cost of removing a normal stroke at the local/global measure. Like experiment 1, most of the existing research work experimented on gestures with a restraint put on pattern variation. This made us perform experiment 2, which would reveal the proposed model’s efficacy in non-ideal conditions. So, we increased the complexity further through the introduction of various patterns for each gesture. In addition to this, 11 new gestures were included along with the gestures of experiment 1. All these gestures are to be gesticulated naturally without any constraint on pattern or speed or scale.
5.3 | Experiment 2: Without reference pattern

For experiment 2, the NITS hand gesture database (for virtual keyboard entry) consisting of 3864 videos of 69 gesture utilised. When it comes to a natural way of gesticulating a gesture in the air, it was observed that the self co-articulations cannot be distinguished by their speed information. In some cases, the velocity of a normal stroke/initial movements (jerk) will be higher compared to the self co-articulations (see Figure 9). Experiment 2 is to show the performance of the proposed model when the gesture has: i) SCA and normal strokes which cannot be distinguished by their speed information; ii) various patterns; and iii) visible scale variation. Figure 9 shows velocity along the x-axis and the y-axis when gesture ‘A’ is performed naturally without any instruction.

When there exists no distinguishable speed information, ‘ED’ and ‘IVe’ failed to capture the self co-articulations; this resulted in gestures removing a part of the first or the last normal stroke in most cases. So, instead of maximum, we used the average velocity, which also resulted in removing a part of self co-articulations and normal stroke in most cases. The MER calculated for ‘ED’ and ‘IVe’ was found to be 0.6514 and 0.6858, respectively. The gesture patterns with no self co-articulations retained its stroke segment coordinates. A higher MER suggests that displacement and velocity fail to identify self co-articulations when gesticulation over the air is done naturally. Based on the movements of the gesture in its initial stage, using (19), ‘Po’ removed the self co-articulations with an MER of 0.0885. Gestures ‘G_4’, ‘J_4’, ‘M_4’, ‘M_2’, ‘N_3’, ‘O_2’, ‘Q_3’, ‘R_3’, ‘W_4’, ‘Z_3’, ‘zero_2’, ‘1_5’, ‘4_3’, ‘4_4’, ‘5_3’, ‘8_4’, ‘@_4’, ‘&_4’, and ‘Top arrow_2’ had its normal stroke segment removed as SCA and normal segments coordinates had the same polarity. Gestures ‘F_6’, ‘division_7’, ‘percentage_7’, ‘question mark_4’, ‘exclamation mark_3’, ‘colon_5’, and ‘semi-colon_4’ had one SCA stroke, while ‘M_6’ and ‘colon_5’ lost two of their normal segments, while ‘E_7’ lost three segments (see the Appendix for Table 2, which provides the information about the loss of segments). The repetition of ‘B_3’ and ‘2_3’ means that two patterns of gesture had the same number of segments. Gestures ‘E_7’, ‘F_6’, ‘J_4’, ‘5_4’, ‘exclamation mark_3’, ‘question mark_3’, ‘colon_3’, ‘semi-colon_3’, ‘colon_5’, and ‘semi-colon_4’ had one SCA stroke, while ‘M_6’ and ‘W_6’ had two SCA stroke segment with them. For the above-mentioned gestures, the selective bypass rules fail to properly bypass the exact SCA segments. For example when ‘J’ and ‘S’ are gestured towards the South, it resulted in four segments with the third segment being SCA stroke, but since it falls under rule 2—South, the second segment will be bypassed. Similarly, ‘question mark_4’ gestured towards the East will have SCA at third segment, but rule 2—East will bypass second segment; ‘M’ gestured towards South will result into six segments, but rule 5—South will bypass third and fifth instead of second and fourth segments; ‘F’ gesture towards South will result in six segments with SCA stroke at second and fourth but rule 4—south will segment third and fifth. These are some of the shortcomings of the formulated bypass rule. The selected coordinates from the local and global measures are fed to the ‘OR’ block, which performs union operation on all the coordinate points. Gestures ‘B_3’, ‘B_3’, ‘G_4’, ‘J_3’, ‘O_2’, ‘Q_3’, ‘zero_2’, ‘1_5’, ‘2_3’, ‘2_3’, ‘4_4’, ‘5_3’, ‘5_4’, and ‘Top arrow_2’ are formed completely because of the availability of missing coordinate points during the ‘OR’ operation (see the Appendix). But gestures ‘J_4’, ‘M_4’, ‘N_3’, ‘R_3’, ‘W_6’, ‘Z_3’, ‘4_3’, ‘8_4’, ‘@_4’, ‘&_4’, and ‘question mark_4’ lost one of its normal segments, while ‘percentage_7’ and ‘division_7’ lost coordinate points of two normal segments (see the Appendix). This is due to the removal of coordinate points from the same segment in both the local and global measures. Gestures ‘E_7’, ‘J_4’, ‘4_4’, ‘5_4’, ‘question mark_3’, ‘colon_3’, ‘semi-colon_4’, ‘colon_5’, ‘percentage_7’, ‘division_7’, ‘question mark_4’, and ‘exclamation mark_3’ had one SCA stroke segment; ‘F_6’, ‘M_6’, and ‘W_6’ had two SCA stroke segments (see the Appendix). Thus, the proposed model for gesture formation had an MER of 0.0371 and 12.03% of self co-articulations.

Analysing the results of the global measure reveals that the system is less dependent on speed information by biased voting. Also, the global measure fails to remove SCA segments, which are aligned with the normal stroke segment. As the distance of coordinates points of SCA stroke segments is larger in gesture gesticulated with high speed, MER of ‘ED’ and ‘IVe’ was lesser in experiment 1 compared to experiment 2. Analysing the results of local measure for both the experiments indicated...
TABLE 2 Failure cases in experiments 1 and 2, segment removed in global and local measure, OR operator output, normal segment lost, and SCA stroke segment

| Gesture failure cases | Segment removed | Segment lost | OR | SCA |
|-----------------------|-----------------|--------------|----|-----|
|                       | Global Local    |              |    |     |
| B_3                   | 0 2 1,2,3       | 0            | 0 0|
| E_7                   | 3,6 2,4,6       | 1,2,3,4,5,7  | 0 3|
| F_6                   | 2 5,5           | 1,2,3,4,5,6  | 0 2,3|
| G_4                   | 2 3             | 1,2,3,4      | 0 0|
| J_3                   | 3 2             | 1,2,3        | 0 0|
| J_4                   | 2,3 2           | 1,3,4        | 2 3|
| M_4                   | 3 3             | 1,2,4        | 3 0|
| M_5                   | 2,3 2,4         | 1,3,4,5      | 0 0|
| M_6                   | 2,4,6 3,5       | 1,2,3,4,5,6  | 0 2,4|
| N_3                   | 3 2             | 1,2,3        | 0 0|
| O_2                   | 2 0             | 1,2          | 0 0|
| Q_3                   | 2 0             | 1,2,3        | 0 0|
| R_3                   | 2 2             | 1,3          | 2 0|
| W_4                   | 3 2             | 1,2,3,4      | 0 0|
| W_6                   | 2,5 3,5         | 1,2,3,4,6    | 5 2,4|
| Z_3                   | 2 2             | 1,3          | 2 0|
| Zero_2                | 2 0             | 1,2          | 0 0|
| 1_5                   | 2 2             | 1,3,4,5      | 0 0|
| 2_3                   | 2 2             | 1,3          | 2 0|
| 4_3                   | 2 2             | 1,3          | 2 0|
| 4_4                   | 3 2             | 1,2,3,4      | 0 3|
| 5_3                   | 3 0             | 1,2,3        | 0 0|
| 5_4                   | 3 2             | 1,2,3,4      | 0 3|
| 8_4                   | 3,4 3           | 1,2,4        | 3 0|
| @_4                   | 2,3 3           | 1,2,4        | 3 0|
| &_4                   | 3,4 3           | 1,2,4        | 3 0|
| Percentage_7          | 2,4,7 2,4,6     | 1,3,5,7      | 4,6 5|
| Division_7            | 2,4,5,7 2,4,6   | 1,3,5,7      | 4,6 5|
| Question mark _4      | 2 2             | 1,3,4        | 2 3|
| Exclamation mark _3   | 0 0             | 1,2,3        | 0 2|
| Semi-colon_4          | 3 2             | 1,2,3,4      | 0 3|
| Colon_5               | 3 2             | 1,2,3,4,5    | 0 3|
| Top arrow_2           | 2 0             | 1,2          | 0 0|

that selective bypass rules fail in gestures with more than two segments having no self co-articulations. These rules were formulated based on the common gesticulation and calligraphy patterns seen in individuals but had failed to remove self co-articulations in some patterns. The union operator assists in overcoming the loss of normal segments to a certain extent. The achieved MER and percentage of self co-articulations in both the experiments exhibit the proposed model performance in terms of total gesture patterns. To analyse the performance of the proposed model for each gesture of the individuals, we calculated the accuracy in terms of correct gesture formed, i.e. the gesture should satisfy: i) GT and ii) no SCA stroke segment. The proposed model in experiment 1 had an overall accuracy of 98.28%, i.e. it failed to remove SCA stroke segments or removed a part of normal segments in 50 gestures of experiment 1. For experiment 2, an overall accuracy of 96.94% was obtained, i.e. 118 gestures did not form the complete gesture or had SCA segments in them. Gesture ‘Z_3’ had the lowest accuracy (53.57%) as most of the individuals in experiment 2 preferred this pattern, and the proposed model incorrectly removed the centre segment. It was observed that out of 118 gestures, 50% of gestures had no self co-articulations. Results of experiments 1 and 2 in terms of MER, percentage of self co-articulations, and accuracy revealed that the proposed model works for both the cases of i) gesticulating with intentional speed variation and ii) gesticulating naturally. The performance measures reveal that the proposed model degrades for gestures (>2 segments) with no self co-articulations and the gesture with strokes aligned in the same polarity.

5.4 Assessment of the proposed model

To evaluate the proposed model with the baseline models, we implemented the work done in [1, 10, 15, 33] on case I: NITS hand gesture database—VI [1] and case-II: NITS hand gesture database (for virtual keyboard entry). The comparison was based on accuracy, computational complexity in terms of Big ‘O’ notation (worst case), and execution time in second (s). Implementation of the existing and proposed work was done in MATLAB with system specification i5-2410M CPU@2.30-GHz processor and 8-GB RAM.

When the existing models were experimented on hand gestures gestured with a reference template (case I), the performance of these models was satisfactory (see Table 3). But when experimented on hand gesture gesticulated without any reference template (case II), the existing baseline models performance got drastically reduced (see Table 3). This confirms the research gap yet to be explored in the field of self co-articulation removal. Hence, there is a need to look for additional characteristics in the gesture to design a system that can

| SCA removal   | Accuracy |
|---------------|----------|
| Case I        | Case II  |
| Jin et al. [33] | 76.21%   | 51.45%   |
| Chen et al. [10] | 84.10%   | 66.20%   |
| Sinha et al. [15] | 96.80%   | 52.73%   |
| Misra et al. [1] | 97.39%   | 53.98%   |

| Proposed | 98.28% | 96.94% |

| Complexity | Time |
|------------|------|
| Case I     | Case II |
| $O(n^3)$  | $O(n^3)$  |
| $O(a^3)$  | $O(a^3)$  |
| $O(\log n)$ | $O(\log n)$ |
| 3.04 s    | 12.45 s   |
provide reasonable performance in both scenarios. In case I, as these gestures had induced speed variations and were gesticulated based on a reference template, the existing models were able to detect the self co-articulations. But the existing models failed to remove self co-articulations in case II since the gestures had more pattern variations and nominal speed differences between the strokes. The drawback of the existing models was the usage of minimum velocity, maximum curvature \([33]\), velocity \((<50\text{ mm/s})\), and position \((<10\text{ mm})\) \([10]\) to detect the self co-articulations. The usage of classifiers to separate the normal and SCA stroke led to the high computational complexity of \(O(n^2)\). The existing model \([1]\) tried to overcome the influence on speed by removing the frames with a velocity less than 50% of average velocity and considered those frames with a velocity greater than average velocity + 50% of average velocity as SCA strokes. This approach even though had less complexity \(O(\log n)\) failed to detect self co-articulations in natural gesticulation over the air. The existing model \([13]\) normalised the speed variations by developing a template matching technique using DTW + Euclidean distance and then removed coordinates with a velocity greater than 50% of average velocity. However, developing a common template for all the patterns of a gesture is not viable, as gestures had patterns that are multi-directional and included self co-articulations at a different position. The usage of template matching technique resulted in the complexity of \(O(n^2)\). Degradation in the accuracy of all the existing models of case II was mainly due to their dependence on speed information to detect self co-articulations and also because of the approach to work only on gestures with a single pattern. On the contrary, the proposed model utilises a biased voting technique that selects only the stroke segments, which satisfies the polarity criteria on the global measure and uses formulated rules to select only the normal stroke segments from the local measure. This makes the system model less dependent on speed information in removing the self co-articulations and robust to variation in pattern and scale of gestures. This is the reason for the proposed system model to perform in par with the baseline models in case I and significantly better in case II. In fact, the proposed model had a relative improvement of 40% over the baseline models and required fewer computations (see Table \(3\)). These results also justify that gestures can be made free from self co-articulations with the help of the proposed model with less dependence on conventional velocity/template matching/machine learning techniques. These results also prove that the proposed system is robust to the changes in pattern/scale/speed of gesticulation compared to the existing works.

Some of the limitations observed in the proposed model while performing these experiments are: a) ‘Po’ in the global measure is sensitive to the change in polarity and hence a threshold of maximal polarity changes, i.e., the gradient polarity should be maximum. Keeping a maximal threshold resulted in the global measure be oblivious to a prior acceptable polarity change, which should have been considered. This made us choose a threshold of minimum 3 for smaller scale and 5 for other scales; b) ‘Po’ of the global measure depends highly on the initial movements of the gesture for choosing either the \(x\)-axis or \(y\)-axis coordinate; hence, it was made sure that the initial movements were exactly correct for the isolated static gesture by removing the jerks at the start and end; c) ‘Po’ is unable to detect a change in polarity when gestures have self co-articulations aligned with normal strokes; d) the trade-off between global/local measure in removing the normal stroke along with SCA stroke was not successful in some cases, as it resulted in removing the normal strokes in gestures with no self co-articulations; and e) the union ‘OR’ operation performed on the outputs of global and local measures helped to amalgamate the normal stroke segments removed by the trade-off, but it puts in the self co-articulations as well.

6 | CONCLUSION

The proposed system model, an integration of global measure (Euclidean distance, instantaneous velocity, and polarity) and local measure (selective bypass) to remove the self co-articulations in isolated dynamic gestures, is developed. Experiment 1—with a reference, 58 gestures with the single pattern, intentional speed variations and fixed scale—had an MER of 0.0069 and 7.40% of the total self co-articulations. Experiment 2—without reference, 69 gestures with patterns variations, natural gesticulation speed, and scale variation—had an MER of 0.0371 and 12.03% of the total self co-articulations. In terms of accuracy, the proposed model had an accuracy of 98.28% and 96.94% in experiments 1 and 2, respectively. The relative improvement of 40% over the baseline models was achieved. Accuracy of existing models in case II and higher MER of ED and IV \(_c\) in experiment 2 proves that speed information cannot solely be used for removing the self co-articulations for natural gesticulation in air. This paper also introduces an ‘mMPo algorithm’ for segmenting the isolated gestures into stroke-level segments, which had an MER of 0.0357. As the performance of the proposed model degrades due to gestures with no self co-articulations, an investigation can be done to check whether pause information exists and can it be utilised to identify these gestures. Also, exploration of supportive characteristics of gestures to distinguish stroke segments, which are aligned in the same polarity, is needed. Improvement of accuracy for failure cases is essential as the accuracy of the pre-processing stage like gesture formation will dictate the performance of the succeeding stages. Identification of the number of SCA strokes may help to design a hierarchical system to develop a gesture classification model. Incorporating semantic segmentation with deep architecture models to remove the SCA strokes can also be explored. Removal of self co-articulations in lowercase English alphabet, continuous hand gestures, and investigating the difference between self co-articulations and co-articulations in continuous gestures would be the future scope of research work.

ACKNOWLEDGEMENT

This work was supported by SERB IMPRINT II: IMP/2018/000098 project [sanction order: SERB/F/10220/2018-2019]. The authors thank Udipta Sarma Bharadwaj, the members of the speech and image processing group, and the student community from the National Institute of Technology
Silchar. The authors gratefully thank the anonymous reviewers for their comments.

**ORCID**

Anish Monsley K. [https://orcid.org/0000-0002-4927-3785](https://orcid.org/0000-0002-4927-3785)

Kaldeep Singh Yadav [https://orcid.org/0000-0002-4007-8502](https://orcid.org/0000-0002-4007-8502)

**REFERENCES**

1. Misra, S., et al.: Vision based hand gesture recognition of alphabets, numbers, arithmetic operators and ASCII characters in order to develop a virtual text-entry interface system. Neural Comput. Appl. 29(8), 117–135 (2018)
2. Misra, S., Laskar, R.H.: Comparative framework for vision based gesturing modes and implementation of robust colour-marker detector for practical environments. IET Image Process. 13(9), 1460–1469 (2019)
3. Singh, J., Laskar, R.H.: ANN-based hand gesture recognition using self co-articulated set of features. IETE J. Res. 61(6), 597–608 (2015)
4. Shamaie, A., et al.: Hand gesture recognition for HCI. [http://www.ercim.eu/publication/Ercim_News/enw46/shamaie.html](http://www.ercim.eu/publication/Ercim_News/enw46/shamaie.html), accessed 29 January 2020.
5. Bhuyan, M.K., et al.: Continuous hand gesture segmentation and co-articulation detection. In: Kalra, P.K., Peleg, S. (eds.). Computer Vision, Graphics and Image Processing. Lecture Notes in Computer Science, vol. 4338, pp. 564–575. Springer New York (2006)
6. Teulings, H.-L., Shoemaker, L.R.B.: Invariant properties between stroke features in handwriting. Acta Psychol. 82(1-3), 69–88 (1993)
7. Viviani, P., Flash, T.: Minimum-jerk, two-thirds power law, and isochrony: Converging approaches to movement planning. J. Exp. Psychol. Hum. Percept. Perform. 21(1), 32–53 (1995)
8. Bhuyan, M.K., et al.: A novel set of features for continuous hand gesture recognition. J. Multimodal User Interface 8(4), 333–343 (2014)
9. Chen, M., et al.: Air-writing recognition—Part I: Modeling and recognition of characters, words, and connecting motions. IEEE Trans. Hum.-Mach. Syst. 46(3), 403–413 (2015)
10. Chen, M., et al.: Air-writing recognition—Part II: Detection and recognition of writing activity in continuous stream of motion data. IEEE Trans. Hum.-Mach. Syst. 46(3), 436–444 (2015)
11. Singh, J., Laskar, R.H.: Self co-articulation detection and trajectory guided recognition for dynamic hand gestures. IET Comput. Vision 10(2), 143–152 (2016)
12. Singh, J., Laskar, R.H.: Recognition of global hand gestures using self co-articulation information and classifier fusion. J. Multimodal User Interface 10(1), 77–93 (2016)
13. Geng, W., et al.: Gesture recognition by instantaneous surface EMG images. Sci. Rep. 6(1), 1–8 (2016)
14. Paul, S., et al.: Handwriting recognition in free space using WIMU-based hand motion analysis. J. Sens. 1–10 (2016)
15. Singh, J., Laskar, R.H.: Hand gesture recognition using two level speed normalization, feature selection and classifier fusion. Multimedia Syst. 23(4), 499–514 (2017)
16. Shin, J, Kim, CM.: Non-touch character input system based on hand tapping gestures using Kinect sensor. IEEE Access 5, 10496–10505 (2017)
17. Kumar, P., et al.: 3D text segmentation and recognition using leap motion. Multimedia Tools Appl. 76(15), 16491–16510 (2017)
18. Viet, V.H., et al.: Spatial-temporal shape and motion features for dynamic hand gesture recognition in depth video. Int. J. Image, Graphics Signal Process. 10, 17–26 (2018)
19. Mukherjee, S., et al.: Fingertip detection and tracking for recognition of air-writing in videos. Expert Syst. Appl. 136, 217–229 (2019)
20. Mantecin, T., et al.: A real-time gesture recognition system using near infrared imagery. PLoS One 14(10), 1–17 (2019)
21. Alam, M.S., et al.: Trajectory-based air writing recognition using deep neural network and depth sensor. Sensors 20(2), 376–392 (2020)
22. Rahim, A., et al.: Hand gesture recognition-based non-touch character writing system on a virtual keyboard. Multimedia Tools Appl. 79(1), 11813–11836 (2020)
23. Yanay, T., Shmueli, E.: Air-writing recognition using smart-hands. Pervasive Mob. Comput. 66, 101183 (2020)
24. Jian, C., et al.: LSTM-based dynamic probability continuous hand gesture trajectory recognition. IET Image Process. 13(12), 2314–2320 (2019)
25. Misra, S., Laskar, R.H.: Development of a hierarchical dynamic keyboard character recognition system using trajectory features and scale-invariant holistic modeling of characters. J. Ambient Intell. Hum. Comput. 10(12), 4901–4923 (2019)
26. Zhai, S., et al.: Foundational issues in touch-screen stroke gesture design—An integrative review. Found. Trends Hum.-Comput. Interact. 5(2), 97–205 (2012)
27. Vikram, S., et al.: Handwriting and gestures in the air, recognizing on the fly. In: Proceedings of Conference of Human Factors in Computing Systems, Paris, France, pp. 1179–1184 (Apr. 2013)
28. Otsu, N.: A threshold selection method from gray level histograms. IEEE Trans. Syst. Man Cybern. 9(1), 62–66 (1979)
29. Rabiner, L.R., Schafer, R.W.: Digital Processing of Speech Signal, 1st ed. Prentice-Hall, Upper Saddle River (1978)
30. Cao, X., Zhai, S.: Modeling human performance of pen stroke gestures. In: Proceedings of the ACM Conference on Human Factors in Computing Systems, pp. 1–12 (Apr. 2007)
31. Willems, D., et al.: iconic and multi-stroke gesture recognition. Pattern Recogn. 42(12), 3303–3312 (2009)
32. Anthony, L., Wobbrock, J.O.: A lightweight multistroke recognizer for user interface prototypes. In: Proceedings of Graphics Interface, Ottawa, ON, Canada, pp. 245–252 (May 2010)
33. Jin, X.J., et al.: Visual gesture character string recognition by classification-basedsegmentation with stroke deletion. In: Proc. 2nd IAPR Asian Conference on Pattern Recognition, Okinawa, Japan, pp. 120–124 (Nov. 2013)
34. Zhou, Y., et al.: Threshold selection and adjustment for online segmentation of one-stroke finger gestures using single tri-axial accelerometer. Multimedia Tools Appl. 74(21), 9387–9406 (2015)
35. Zheng, H., et al.: Chinese/English mixed character segmentation as semantic segmentation. arXiv:1611.01982v2 (2016)
36. Feitosa, R.A., et al.: Multidimensional representations for the gesture phase segmentation problem—An exploratory study using multilayer perception. In: Proc. 10th International Conference on Agents and Artificial Intelligence, Funchal, Portugal, pp. 347–354 (Jan. 2018)

**APPENDIX**

Table 2 elucidate the working of union OR operator to form the complete gesture. It is to be noted that union ‘OR’ operation is performed at coordinate level and just to visualise the operations we have referred as segments (see Table 2).