Human Computer Interaction Based Hand Gestures Using Deep Learning and Region (ROI) of Interest Algorithm

Mr. Arivarasan S, Assistant Professor ,
Department of Information Technology,
V.S.B. Engineering College,
Karur, Tamilnadu, India.
E-mail
:arivarasan.vsbengineering2021@gmail.com

Mr. Arunkumar S, UG Scholar,
Department of Information Technology,
V.S.B. Engineering College,
Karur, Tamilnadu, India.
E-mail: arunsak6002@gmail.com.

Mr. Deepak M, UG Scholar,
Department of Information Technology,
V.S.B. Engineering College,
Karur, Tamilnadu, India.
E-mail: deepakmohanraj8@gmail.com.

Mr. Manoj R, UG Scholar,
Department of Information Technology,
V.S.B. Engineering College,
Karur, Tamilnadu, India.
E-mail: manojr21599@gmail.com

ABSTRACT:

Using the hand directly as an input device is attractive method for providing natural human-computer interactions (HCI). Each display system captures a certain amount of spatial information (such as spatial) upper body or movement of one hand), and the whole system works two temporary scales. We provide a review of vision-based hand gesture recognition algorithms. We point out the need to consider these steps together with the authenticity accuracy of the algorithm predict its success in real world applications. Starting from this pixel, a maneuvering search algorithm allows for the identification of the entire hand outline. However, this technology has many drawbacks that prevent simplicity and naturalness this allows the user to interact with the computer control environment, which requires lengthy calibration and setup procedures. Computer vision has the potential to provide much more natural, non-contact solutions. As a result, there have been considerable research efforts to use the hand as an input device for HCI. Successful efforts in hand gesture recognition research within the last two decades paved the path for natural human-computer interaction systems. Unresolved challenges such as reliable identification of gesturing phase, sensitivity to size, shape, and speed variations, and issues due to occlusion keep hand gesture recognition research still very active. Methods for integrating identity classification results into the full SLR are in development further adaptations towards speech recognition techniques sign in the specific case. Finally current boundaries and recent research are discussed given. It involves incessant identity recognition and work towards the real signer is freedom, how to effectively combine different methods suitable for identification, use of current linguistic research and high volume data sets.

INTRODUCTION

More emphasis has recently been placed on HCI (Human Computer Communication) research to facilitate the use of interfaces and the direct use of natural communication the manipulative ability of humans. Skill learning many advanced applications that require systems, surgical simulations, and general robot instruction or direct sensitivity of hand and / or finger movements to virtual environments. Gestures are present in most everyday human actions or activities, and participate in human interactions in situations that require quiet communication (underwater, quiet situations, secret communication, etc.) by completing speech or as an alternative to spoken language. etc.) or for the hearing impaired. An important aspect of our approach is the use of a multi-modal neural network to characterize so-called dynamic poses of different periods (i.e. temporal scales). We use different data channels to distort each gesture not only temporarily but also spatially on many levels, providing the environment for body movement over the body and more elegant hand / finger expression. Key components of the hand recognition system are: data acquisition, hand distribution (e.g. division and surveillance), hand feature recognition and sign recognition based on identified features. Different approaches can be used for hand normalization in the purchased data, depending on its nature. In terms of in-depth data, the classical solution is empirically or automatically in-depth input. Experienced solutions select the limits of the most probable search space by trial and error and concentrate the computational effort for manual localization within it.

METHODOLOGY:

We formulate a pose descriptor, consisting of 7 logical subsets, and allow the classifier to perform online feature selection. The descriptor is calculated based on 11 upper body joints, relevant to the task, whose raw, i.e. pre-normalization, positions in a 3D coordinate system associated with the depth sensor are denoted as $p^{(i)}_{raw} = \{x^{(i)}, y^{(i)}, z^{(i)}\}$, i = 0...10 (i = 0 corresponds to the HipCenter joint).

Following the procedure proposed in [38], we first calculate normalized joint positions, as well as their velocities and accelerations, and then augment the descriptor with a set of characteristic angles and pairwise distances.

Joint positions: The skeleton is represented as a tree structure with the HipCenter joint playing the role of a root node. Its coordinates are subtracted from the rest of the vectors p_{raw} to eliminate the influence of position of the body in space. To compensate for differences in body sizes, proportions and shapes, we start from the top of the tree and iteratively normalize each skeleton segment to a corresponding average "bone" length estimated from all available training data. It is done in the way that absolute joint positions are corrected while corresponding orientations remain unchanged:

$$p^{(i)}(t) = p^{(i)}_{raw}(t) + \frac{p^{(i)}_{raw}(t) - p^{(i-1)}_{raw}(t)}{\left\| p^{(i)}_{raw}(t) - p^{(i-1)}_{raw}(t) \right\|} b^{(i=1,\,i)} - p^{(0)}_{raw}(t), \quad (1)$$

where $P_{raw}^{(i)}$ is a current joint, $P_{raw}^{(i-1)}$ is its direct predecessor in the tree, $b^{(i-1,i)}$, i = 1 . . . 10 is a set of estimated average lengths of "bones" and p are corresponding normalized joints. Once the normalized joint positions are obtained, we perform Gaussian smoothing along the temporal dimension (σ = 1, filter size 5 × 1) to decrease the influence of skeleton jitter. Joint velocities are calculated as first derivatives of normalized joint positions: $\delta p^{(i)}(t) \approx p^{(i)}(t+1) - p^{(i)}(t-1)$. Joint accelerations correspond to the second derivatives of the same positions: $\delta^2 p^{(i)}(t) \approx p^{(i)}(t+2) + p^{(i)}(t-2) - 2p^{(i)}(t)$. Inclination angles are formed by all triples of anatomically connected joints (i, j, k), plus two "virtual" angles (Right,Left)Elbow-(Right,Left)HandHipCenter.

$$\alpha^{(i,j,k)} = \arccos \frac{(p^{(k)} - p^{(j)})(p^{(i)} - p^{(j)})}{\left\|p^{(k)} - p^{(j)}\right\| \cdot \left\|p^{(i)} - p^{(j)}\right\|} \quad (2)$$

Azimuth angles β provide additional information about the pose in the coordinate space associated with the body. We apply PCA on the positions of 6 torso joints (HipCenter, HipLeft, HipRight, ShoulderCenter, ShoulderLeft, ShoulderRight) to obtain 3 vectors forming the basis: $\{u_x, u_y, u_z\}$, where u_x is approximately parallel to the shoulder line, u_y is aligned with the spine and u_z is perpendicular to the torso. Then for each pair of connected bones, β are angles between projections of the second bone (v_2) and the vector u_x (v_1) on the plane perpendicular to the orientation of the first bone. As in the previous case of inclination angles, we also include two virtual "bones" (Right,Left)Hand-HipCenter.

$$v_1 = u_x - (p^{(j)} - p^{(i)}) \frac{u_x \cdot (p^{(j)} - p^{(i)})}{\|p^{(j)} - p^{(i)}\|^2}$$

$$v_2 = (p^{(k)} - p^{(j)}) - (p^{(j)} - p^{(i)}) \frac{(p^{(k)} - p^{(j)}) \cdot (p^{(j)} - p^{(i)})}{\|p^{(j)} - p^{(i)}\|^2} \quad (3)$$

$$\beta^{(i)} = \arccos \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$$

Bending angles γ are a set of angles between a basis vector u_z, perpendicular to the torso, and normalized joint positions:

$$\gamma^{(i)} = \arccos \frac{u_z \cdot p^{(i)}}{\|p^{(i)}\|} \quad (4)$$

Pairwise distances. Finally, we calculate pairwise distances between all normalized joint positions:

$\rho^{(i,j)} = \| p^{(i)}_n - p^{(j)}_n \|$. Combined together, this produces a 183-dimensional pose descriptor for each video frame: $D = [p, \delta p, \delta^2 p, \alpha, \beta, \gamma, \rho]^T$. Finally, each feature is normalized to zero mean and unit variance. A set of consequent 5 frame descriptors sampled at a given step s are concatenated to form a 915-dimensional dynamic pose descriptor which is further used for gesture classification. The two subsets of features involving derivatives contain dynamic information and for dense sampling may be partially redundant as several occurrences of same frames are stacked when a dynamic pose descriptor is formulated. Although theoretically unnecessary, this is beneficial in the context of a limited amount of training data.

ALGORITHM:

Computation of intensity statistics for ROIs involves multiplying pixel intensities by the area of pixels that lie with the outline of the ROI. ... If the row intersects the ROI anywhere, then each pixel in that

image row is considered in turn. Each pixel is intersected with the ROI. Spline and Elliptical ROIs present a particular issue, since calculation of intensity statistics with a truly curved outline would be problematic and very slow. Calculation of intensity statistics for Spline and Elliptical ROIs is therefore done by representing the outline by a series of straight line segments.

Algorithm: Counter Point Detection

Input:

The first point, P1, of a list of contour points belonging to a potential hand contour, C

Output:

List of contour points Pi belonging to an identified contour C

For each potential hand contour C

From the first point P1 found, trace in order each point of contour and stop when the C.length > 700 or when the next point is already present in the contour list.

//if the last point is near the first point

if ((C.length > 10 and (|LastPoint.x – P1.x| <= 15 pixels and |LastPoint.y - P1.y| <= 15 pixels and |LastPoint.z- P1.z| <= 15 pixels)) then the Contour C is accepted Display the contour points in red

else

the Contour C is rejected

end

The gain in performance and speed comes from multiple improvements in the various stages of the solution, including a faster algorithm to search the first pixel of the hand contour; the recognition of gestures independent of their position in the field of view of the Kinect; the use and adaptation of DTW not only as a validation tool.

Raw data coming from the Kinect is used to recuperate depth information on all the pixels of an image. A novel, faster algorithm is then proposed to identify each point of a closed contour identified within a given depth interval.

Starting from the recuperated contour points, the center of the palm is identified as the center of the largest circle circumscribed in the contour. The fingertips are localized by employing the k-curvature algorithm, which is based on the change in the slope angle of the tangent line at selected points over the contour.

It aims to develop algorithms and methods to correctly identify a sequence of produced signs and to understand their meaning. Many approaches to SLR incorrectly treat the problem as Gesture Recognition (GR). So research has thus far focused on identifying optimal features and

classification methods to correctly label a given sign from a set of possible signs. However, sign language is far more than just a collection of well specified gestures.

Much effort has also been put into spatiotemporal invariant features. In (Yuan et al., 2011) the authors propose a RGB action recognition system based on a pattern matching approach, named naive Bayes mutual information maximization (NBMIM). Each action is characterized by a collection of spatiotemporal invariant features which are matched with an action class by measuring the mutual information between them. Based on this matching criterion, action detection is to localize a subvolume in the volumetric video space that has the maximum mutual information toward a specific action class. A novel

spatiotemporal branch-and-bound (STBB) search algorithm is designed to efficiently find the optimal solution. Results show high recognition results on KTH, CMU, and MSR data sets, showing speed up inference in comparison with standard 3D branch-and-bound.

Data modalities integrated by our algorithm include intensity and depth video, as well as articulated pose information extracted from depth maps. We make use of different data channels to decompose each gesture at multiple scales not only temporally, but also spatially, to provide context for upper-body body motion and more fine-grained hand/finger articulation.

EXISTING SYSTEM:

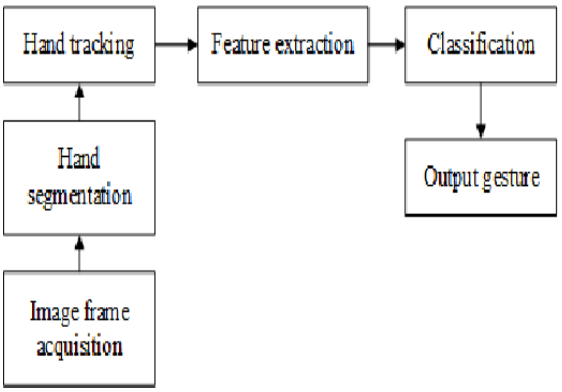
Gesture Recognition consists of two approaches a) vision based b) glove based. Glove based approach uses sensors or gloves to identify the hand gesture. Some type of flex sensors, accelerometers etc are used in glove based approach. Static gestures use hand poses and the image is captured by using cameras.The images captured are given for analysis which is done using segmentation. A drawback of 3D reconstruction is the additional computational cost. However, 3D information is valuable data that can help eliminate problems due to self-occlusions which are inherent in image-based approaches.

In addition to the main pipeline, we have created a baseline model based on an ensemble classifier trained in a similar iterative fashion but on purely handcrafted descriptors. It was done to explore relative advantages (and disadvantages) of using learned representations and also the nuances of fusion. In addition, due to differences in feature formulation as well as in the nature of classifiers, we found it beneficial to combine the proposed deep network with the baseline method in a hybrid model as separately two models make different errors. Due to the relative disadvantages of HMMs (poor performance when training data is insufficient, no method to weight features dynamically and violations of the stochastic independence assumptions), they coupled the HMM recogniser with motion analysis using computer vision techniques to improve combined recognition rates.

HMM based methods are effective and are widely used for HGR. However HMM based approaches require a large number of training samples and have the disadvantage of elaborate training procedure. The computational costs of HMM based algorithms increase with the gesture vocabulary. In addition, the performance of HMM based algorithms reduces when there are variations between training and testing conditions. Finding the optimal parameter sets and trajectory spotting for temporal segmentation are other bottlenecks in using HMM.

The MMI term refers to the capture and conversion of signals related to the appearance, behavior, or physiology of a human via a computer system. All interface techniques that rely on sound or vision have contributed to a radical change in the mechanism of operating a computer.

EXISTING BLOCK DIAGRAM



PROPOSED SYSTEM

The hand gesture images are captured from the vision based camera. Using background subtraction technique to separate the hand from background. Segmentation and classification technique to classify the finger counts. Open application interfaces based on finger count. Voice based alert system.

Hand image acquisition:

Hand image captured from web camera. The purpose of Web camera is to capture the human generated hand gesture and store its image in memory. The package called .NET framework is used for storing image in memory. Digital Image Processing
in image processing, it is defined as the action of retrieving an image from some source, usually a hardware-based source for processing. It is the first step in the workflow sequence because, without an image, no processing is possible. Palm vein recognition systems, like many other biometric technologies, capture an image of a target, acquire and process image data and compare it to a stored record for that individual.

Background Subtraction:

Extract the foreground from background image. Using Binarization approach to assign the values to background and foreground. Foreground pixels are identified in real time environments. Edge based features require very simple background to be effective. Distance transforms of edges help to calculate more robust error measures. Chamfer matching was used in several studies demonstrating good performance in cluttered backgrounds.

ROI extraction:

A region of interest (often abbreviated ROI), are samples within a data set identified for a particular purpose. The concept of a ROI is commonly used in many application areas. Hand regions are extracted and provide segmented results. For example, in medical imaging, the boundaries of a tumor may be defined on an image or in a volume, for the purpose of measuring its size. High-resolution computed tomography (CT) reconstructions currently require either full field of view (FOV) exposure, resulting in high dose, or region of interest (ROI) exposure, resulting in artifacts.

Finger count detection:

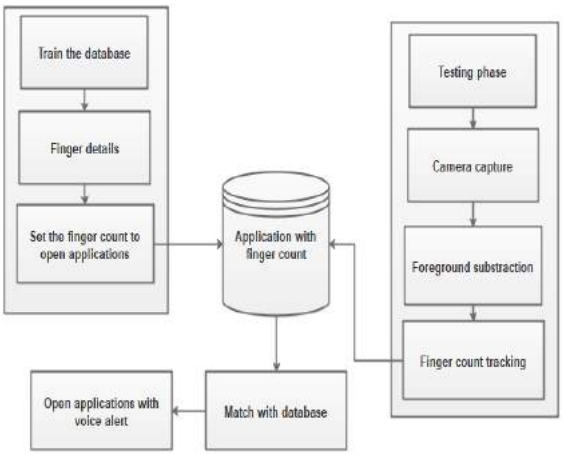
Skin can be easily detected by using the color information. Edge detection is applied to separate the arm region from

the hand region. The hand gestures information consists of skin color, movement and edge feature. This is a system which detects a human hand, segments the hand using thresholding, counts the number of fingers being held up and displays the finger count, all from a live video input.

Application creation:

Based on finger count, application can be open. Using training side, we can save the application which opened by finger count. Provide voice alert at the time of application open. We begin training by presenting modality-specific parts of the network with samples where only one modality is present. In this way, we pre-train initial sets of modality-specific layers that extract features from each data channel and create more meaningful and compact data representations. From the parameters of the fitted surface model at each stage, a characteristic feature vector was created, when combined with Radial Basis Function Interpolation networks it can be used to accurately predict the pan, tilt and roll of the head. The hand creates images that are very difficult to analyze in general. High level features such as fingertips, fingers, joint locations, and the links between joints are are very desirable but also very difficult to extract in a bottom-up manner. The algorithms that require direct extraction of high level features often rely on markers to extract fingertip joint locations or some anchor points on the palm [18, 5, 11]. Assuming a clutter-free background, it is possible to extract some high level features without any markers.

BLOCK DIAGRAM



This approach is simple and does not impose significant limitations for the user, as we naturally tend to keep the hands in front of body when we gesticulate. In order not to have to be constrained by the closest pixel, a possible solution is to use other elements in the Kinect image as a reference, such as using of the skeleton tracking feature of Kinect to identify the hand position. However, due to the fact that this joint is among the least stable joints to track using the Kinect SDK [38], this solution was not retained in this work.

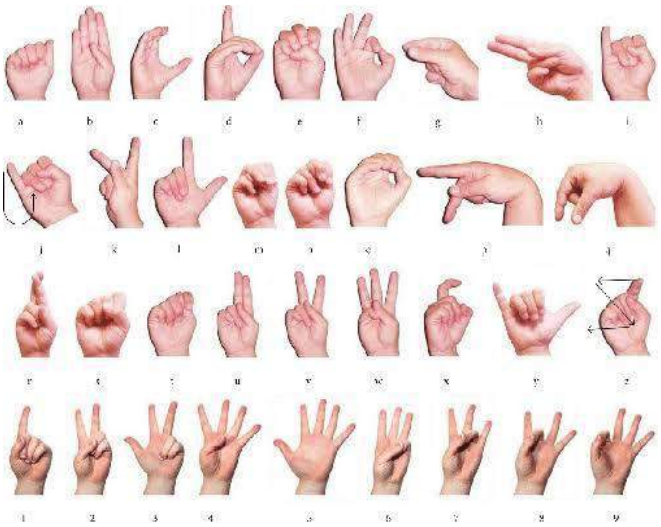
Similar to the approach of [6], the hand palm center is calculated using the contour pixels as well as the interior pixels of the hand. To identify the bounding box of the hand, the minimum and maximum point on the X and Y axis are computed. The minimal distance (dmin) of interior pixels is then calculated for each block of 5×5 pixels with respect to each contour pixel. One

contour pixel in 5 was proved experimentally to be sufficient for the final estimation of the hand center. The interior pixel that has the maximal value for dmin is considered to be the center of the palm. In simple terms, the center of the palm is the center of the largest circle that can be circumscribed in the palm, as illustrated in Fig. 4c for the case of the OK sign.

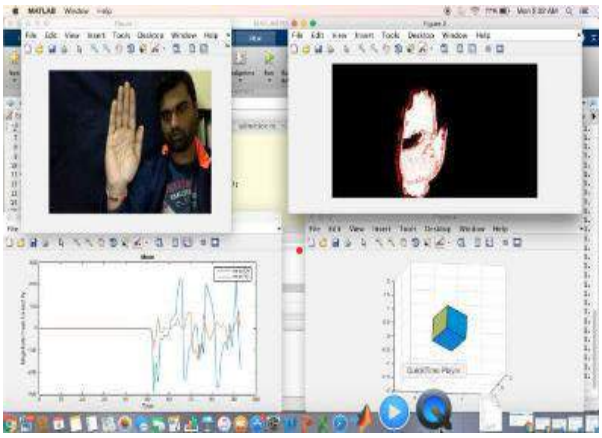
During the final phase, labels for validation data were published and the participants performed similar tasks as those performed in previous phase, using the validation data and training data sets in order to train their system with more gesture instances. The participants had only few days to train their systems and upload them. The organizers used the final evaluation data in order to generate the predictions and obtain the final score and rank for each team. At the end, the final evaluation data was revealed, and authors submitted their own predictions and fact sheets to the platform.

IMPLEMENTATION

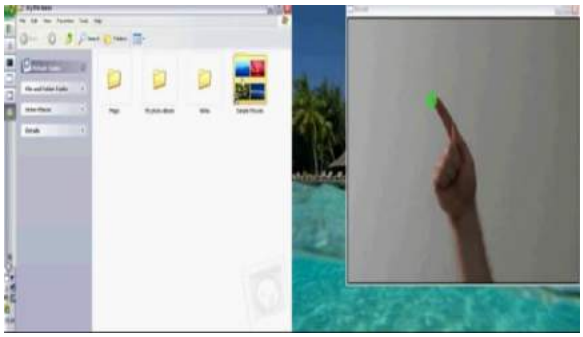
Hand location, angle and velocity features are combined in [11] to implement an HMM for HGR. Hand is localized by skin-color analysis and tracked by connecting the centroid of moving hand regions. The paper compared the utility of the three features, location, angle, and velocity, and concluded that angular features are most effective, having better discriminative power.



Hand gesture is a very natural form of **human interaction** and can be used effectively in **human computer interaction** (**HCI**). This project involves the design and **implementation** of a **HCI using** a small **hand**-worn wireless module **with** a 3-axis accelerometer as the motion sensor.



This is an interesting and important concept as it suggests that when the signer's mouth is occluded it is not necessary to know the mouth shape. Instead they believe that it can be inferred by the information at either side, in a similar manner to a human observer. While the theory is correct, the implementation may prove challenging.



The fingertip-based algorithms described above can be seen as a special case of this approach. Instead of fingertips, they used rotation and scale invariant moments of the hand silhouette. The mapping was implemented using a machine learning architecture (Specialized Mapping Architecture (SMA)) without applying any extra hand motion constraints. SMA is capable of generating multiple hypotheses.

CHALLENGE RESULTS

The challenge attracted high level of participation, with a total of 54 teams and near 300 total numbers of entries. This is a good level of participation for a computer vision challenge requiring very specialized skills. Finally, 17 teams successfully submitted their prediction in final test set, while providing also their code for verification and summarizing their method by means of a fact sheet questionnaire. After verifying the codes and results of the participants, the final scores of the top rank participants on both validation and test sets were made public: these results are shown in Table 6, where winner results on the final test set are printed in bold. In the end, the final error rate on the test data set was around 12%.

Feature Extraction and Matching

The hand creates images that are very difficult to analyze in general. High level features such as fingertips, fingers, joint locations, and the links between joints are are very desirable but also very difficult to extract in a bottom-up manner. The algorithms that require direct extraction of high level features often rely on markers to extract fingertip joint locations or some anchor points on the palm [18, 5, 11]. Assuming a clutter-free background, it is possible to extract some high level features without any markers. [27] uses a markerless fingertip extraction algorithm based on Gabor filters and a special neural network architecture (LLM-net). In [24], contour analysis was performed to detect the intersections of the fingers and the palm.

The majority of studies rely on low-level features that are utilized in matching error calculations during the model fitting phase. The calculation of the matching error requires: (1) extracting a set of features from the input images, (2) projecting the model on the scene (or back-projecting the image features in 3D), and (3) establishing a correspondence between groups of model and image features. It can also be argued that 3D pose data can provide more useful features for gesture recognition purposes as they are view independent and directly related to the hand motion.

CONCLUSION:

In this paper, we have reviewed a number of studies addressing the

problem of full hand motion estimation. The existence of an expensive but high speed system is quite encouraging [35]. However, the lack of an implementation that is part of a real world system indicates that there is still a lot of open theoretical questions.

It can elegantly cope with more spatial or temporal scales. Beyond scaling, an interesting direction for future work is a deeper exploration into the dynamics of cross-modality dependencies.

While the community continues to discuss the need for including non-manual features, few have actually done so. Those which have [2, 5], concentrate solely on the facial expressions of sign. There is still much to be explored in the veins of body posture or placement and classifier (hand shape) combinations.

The general trend in these systems is building single camera systems. However, there is also an inevitable tendency to avoid occlusions by keeping the global hand pose fixed with respect to the camera. Multiple camera systems and 3D features are not explored very well. Although these systems are more expensive, they can provide better ways to handle occlusions and can lead to more accurate hand tracking systems for advanced tasks such as virtual object manipulation.

REFERENCES:

H. Cooper, B. Holt, and R. Bowden, ‘‘Sign language recognition,’’ in Visual Analysis of Humans. Berlin, Germany: Springer, 2011, pp. 539–562.

S. Escalera, V. Athitsos, and I. Guyon, ‘‘Challenges in multi-modal gesture recognition,’’ in Gesture Recognition. Berlin, Germany: Springer, 2017, pp. 1–60.

N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, ‘‘Multi-scale deep learning for gesture detection and localization,’’ in Proc. Eur. Conf. Comput. Vis. Berlin, Germany: Springer, 2014, pp. 474–490.

G. Plouffe and A.-M. Cretu, ‘‘Static and dynamic hand gesture recognition in depth data using dynamic time warping,’’ IEEE Trans. Instrum. Meas., vol. 65, no. 2, pp. 305–316, Feb. 2016.

A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, ‘‘Visionbased hand pose estimation: A review,’’ Comput. Vis. Image Understand., vol. 108, nos. 1–2, pp. 52–73, Oct. 2007.

Manar D. Samad, IEEE “A Feasibility Study of Autism Behavioral Markers in Spontaneous Facial, Visual, and Hand Movement Response Data ” 2018.

Biswas, Amrita “Finger Detection for Hand Gesture Recognition Using Circular Hough Transform” 2018.

Devineau, Guillaume “Deep Learning for Hand Gesture Recognition on

Skeletal Data" IEEE International Conference on. IEEE, 2018

Sahoo, Jaya, Samit Ari "Hand Gesture Recognition using DWT and F-ratio Based Feature Descriptor." IET 2018