American Sign Language (ASL) Recognition

Demo available @

Saikat RoyChowdhury (rychwdh2@illinois.edu)
Shyam Rajendran (srajend2@illinois.edu)
Udit Mehrotra (umehrot2@illinois.edu)

Masters Computer Science,
University of Illinois, Urbana Champaign

Abstract

This work is part of a vision based American Sign Language Interpretation (ASL) system for natural human computer interface. It comes under the umbrella of much broader research field of hand gesture recognition. The aim of the paper is to develop a robust system which can interpret the American sign language digits from the static images, as well as video stream inputs. The system involves various important components like Hand segmentation from the background, extracting features of the hand and a classifier for predicting the sign given the features. We also developed an algorithm for extracting the hand sign frames from a video input, which can then be used for classification.

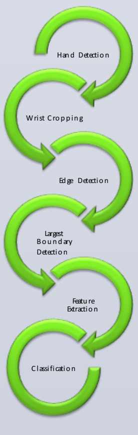
Prior research

There are two major approaches for Hand gesture recognition: Data glove and Vision based. The Data glove approach uses a glove with sensors attached to track finger blending, positioning and orientation. In this paper we have gone ahead with vision based approach, as it is more cost effective and feasible since the user does not have to wear a cumbersome device. In hand based gesture recognition systems, hand tracking and segmentation are the most important and challenging steps towards gesture recognition. Uncontrolled environment, lighting conditions, noisy background, skin color detection and rapid hand movements are some of the challenges that need to be considered while capturing and tracking the hand gestures.

The second major step after Hand segmentation is Shape Feature extraction from the segmented hand. The shape feature extraction can usually be divided into contour based and region based. In our work, we have used contour based method which extracts shape feature information from the boundary of the entity. We have used Fourier descriptor method as it can over come the effect of noise and boundary variations by analyzing the shape in spectral domain. Furthermore, they are compact, computationally light and there matching is a relatively simple process.

The final step of the process involves training a classifier using the shape features extracted from the training data set collected, and using it to predict or classify a given gesture. In our work we have tried various methods like Neural Networks, K nearest neighbor and PCA for the purpose of classification, and have included the results obtained from each of the methods.

Algorithm


Hand Detection
Wrist Cropping
Edge Detection
Largest Boundary Detection
Feature Extraction
Classification

Execution Methodology

1. Hand Detection:

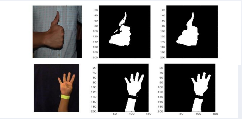
- Primary task skin color detection . A pixel is classified as a skin pixel if it satisfies an OR condition of the following formulas in RGB space and YCbCr space.

$$R > 95 \;\&\; G > 40 \;\&\; B > 20$$
$$\text{Max}\{R,G,B\} - \text{Min}\{R,G,B\} > 15 \;\&$$
$$|R - G| > 15 \;\&\; R > G \;\&\; R > B$$

RGB based formulae

$$Y = 0.299R + 0.587G + 0.114B$$
$$Cr = 128 + 0.5R - 0.418G - 0.081B$$
$$85 < Cb < 135 \;\&\; 135 < Cr < 180 \;\&\; Y > 80$$

YCbCr based formulae



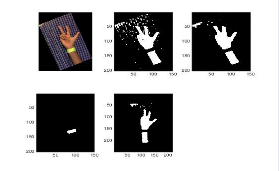
Median Filter & Morphological Close operation

2. Smoothen segmentation output:

- 2-D median filtering technique to reduce segmentation output where each pixel is the median of n X n neighborhood. (n = 4,7)
- Dilation and Erosion methods used for morphological close operation.

3. Wrist Cropping

- Computationally intensive with hand tilt.
- PCA application on entire hand was not sufficient.
- PCA on wristband helped to find principle component orientation making hand rotation invariant analysis.
- Wrist cropping is significantly less computation intensive if the forearm and wrist orientation is known. A simple width based wrist detection algorithm can then be applied.



From right top left to bottom in clockwise (1) Original tilted hand (2) skin color based segmentation (3) image smoothing using median filter/morphological close (4) Wrist band Detection and PCA for orientation detection (5) Rotation normalized image.

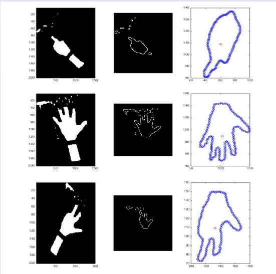
4. Edge Detection (Canny Edge detection)

- Gaussian filter to smoothen noise
- Find intensity gradients of the image
- Apply non-maximum suppression to get rid of spurious response
- Apply double threshold to determine potential edges
- Use a hysteresis based approach to track edges and suppress weak edges

5. Largest Boundary Detection

(Moore-Neighbourhood tracing algorithm)

- To detect the boundaries and select the biggest boundary to separate objects that had similar skin tone color.
- Assumption that the palm and fingers form the biggest boundary, which is true most of the time. Such a method helps us in eliminating skin color background objects (like a distant face etc).



From top to bottom (1) segmented hand based on skin color (2) edge detector output (3) Select the largest boundary using Moore-Neighbourhood Algorithm and arc-tan sampled points of the boundary

6. Feature Extraction

- Used Contour based feature extraction method through use of Fourier descriptors.
- Computed shape signature functions which can then be analyzed in the frequency domain.
- **Equal Points Sampling** and **Equal Arc Length Sampling** were tried and **Arc length** sampling finalized due to better results.
- Two Shape signature functions were tried: **Centroid distance** and **Complex coordinates.**

Complex coordinates:

- Complex number generated from boundary coordinates

$$z(t) = x(t) + iy(t)$$

- To eliminate the effect of bias, the coordinates are shifted by subtracting the centroid i.e

$$z(t) = [x(t) - x\_c] + i[y(t) - y\_c]$$

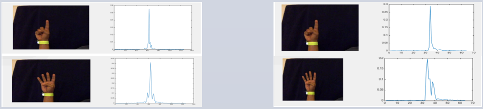
where (x_c, y_c) is the centroid of the shape. This shift makes the shape invariant to translation.

Centroid distance:

- It is expressed as the distance of the boundary points, from the centroid (x_c, y_c) of the shape

$$r(t) = ([x(t) - x\_c]^{2} + [y(t) - y\_c]^{2})^{1/2}$$

- **Discrete Fourier Transform** on the 64 point normalized shape signature to obtain **Fourier Descriptors**, is rotation invariant considering only magnitude and normalized in scale.

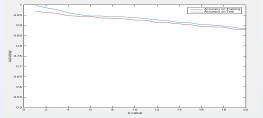


Fourier Descriptor plots for each of the few sample digits for a sample training set using both Complex coordinates (left) and Centroid distance (right) as the shape signatures.

7. Classification

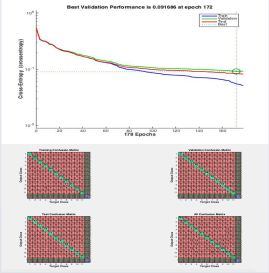
K Nearest Neighbor based Classification:

- 3800 samples of ASL digits,
- training : testing data = 70% : 30%
- ~ 95% accuracy when k = 1;
- Average result obtained around 55 - 65% on new test samples.



Neural Network based Supervised Learning:

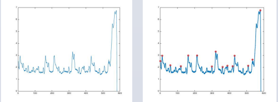
- Feed forward neural network with sigmoid transfer function
- Back propagation algorithm for learning weights and biases for neurons.
- Utilized 64 point Fourier Descriptors features per digit with N' hidden neurons and 11 output neurons (digits 0 to 10) : N=48 with experimentation for reliable accuracy
- Average result on 3800 samples: 84% on the testing set, 91% on the training set and 83% on the validation.
- 70-80% accuracy on random set on average.



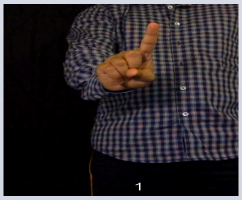
From top to bottom (1) EXPLAIN the plots

8. Video frame gesture detection

- XOR and frame difference technique tried.
- Peak finding algorithm gives us the frames that constitute the actual gestures.



From left to right (1) plot generated using adjacent frame difference. (2) peak detector output.



Below is a screen grab of video subtitle generated by taking mode of the predicted classes with in a set window of images.