K-016

The Effect of Size and Position Normalization on HMM-based JSL Recognition

Elghadi Abdulhakim Asai Shingo Sako Shinji Kitamura Tadashi

Abstract

This paper shows a technique of continuous Japanese Sign Language (JSL) recognition based on Hidden Markov models (HMMs). The system aims for an automatic signer independent recognition of JSL sentences. Stereo video camera is utilized for extracting geometric properties of the head and hands trajectory. Size and position normalization is employed in order to account for differences in body-size and proportions. The system was implemented and evaluated using the RWC JSL database consisting of 64 sentences and 4 signers. The recognition results are then compared to the conclusion that normalization indeed increases the recognition rate of the system.

1. Introduction

Sign language is a language that relies on body – especially limb– movements signify words. Each word has a movement assigned to it. This movement can be used in the context of sign language recognition for training the recognition system and later recognizing new words and whole sentences. One way to achieve this is by extracting the coordinates of the track for each hand movement and inputting them as feature vectors to the recognition system in the learning and recognition phase[2]. The merit of this method is that it does not require complex high-level feature extraction e.g phonetic features. On the other hand the reliance on hand coordinates tracks as feature vectors introduces with it differences in tracks induced by differences in body size and relative position of the signer. These differences introduce discrepancies between tracks of the same sign made by different persons. In this paper we try to investigate the effectiveness of correcting such discrepancies using position normalization and size normalization. We start by describing the database used. Then we explain the method we used for data extraction in section 2.1. In section 2.2 we outline the normalization method we used. And finally we compare the results we obtained from our experiment in section 2.3. Section 3 concludes the paper with some remarks.

2. Experiment

The data we used is from the RWC database of JSL[1]. The database contains the data of 4 signers 2 females and 2 males. Each signer signs the same set of 64 sentences composed from 41 word vocabulary set, and each signer is filmed in stereo using a left view camera and a right view camera. Finally each recording is done twice for each signer. Table 1 shows more detailed information about the database.

2.1 Hand coordinates extraction

We reduced the 2^{24} colors space of the images to a more manageable 64 color space using a 64 cluster using k-means algorithm. We then manually selected the colors closest to skin color and masked each frame to remove every other color that does not belong to the skin color set. This produced frames with three skin patches one for the face and two for the left and right hands. To find the centroids of each hand we used a 3 cluster k-means algorithm with initial centroids close to the hands and face for the first frame and with the previous frame centroids as the initial centroids for the other frames.

Since the k-means algorithm occasionally wanders astray when trying determining the desired centroids, we used a correction procedure based on the position of the previous frame's centroids and the current frame's red wrist bands centroids which are much more well-behaved with respect to centroid calculation, and calculated the expected displacement of the x and y directions. See figure 1. This correction method can be summarized as follows.

- Calculate wrist displacement for frame f: $dx^f_{wrist} = x^f_{wrist} - x^{f-1}_{wrist}$
- Calculate displacement threshold for frame f: $x^f_{thresh} = dx^f_{wrist} \times \alpha$ Where α has the value of 1.5 which is empirically determined‡.
- if $x^f_{hand} - x^{f-1}_{hand} \notin [0, x^f_{thresh}]$ then $x^f_{hand} \leftarrow x^{f-1} + dx^f_{wrist} \times \beta$ where β is determined empirically to be 1.15. As hands tend to make slightly larger displacements from one frame to the next β serves to magnify wrist movement. See figure 1.

Table 1: RWC Gesture Database

Number of signers	4 persons (2 females, 2 males)
Sentences	64 JSL sentences×2 recordings×2 views (stereo)
Image size	Horizontal 320×Vertical 240 pixels
Color depth	24 bit/pixel(RGB: 8bit each)
Frame rate	30 frames/sec
Number of frames per sentence	121∼160

2.2 Normalization

We normalized for both position and size in this experiment. For position, we considered the face centroid to be the origin and shifted all hand track coordinates accordingly. $x^f_p = x^f - x^1_{face}$ where superscripts are frame numbers, and x_p is the x coordinate after position normalization. We did the same calculation for y-axis position normalization.

For horizontal size normalization, we calculated the horizontal distance between the white shoulder marks for each frame and then calculated the normalization

†Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology, NIT

‡Because hand's shape is oblong. The centroid should be allowed tolerance in terms of position

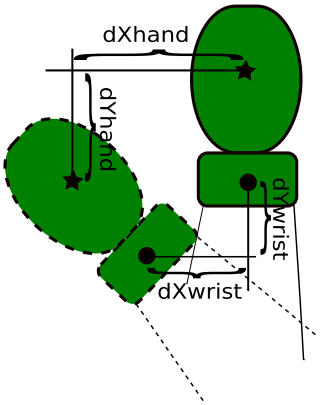

Figure 1: Comparison between hand movement and wrist movement

Figure 2: Initial frame centroids

factor for each frame as $\frac{\Delta x_{av}}{\Delta x^s}$ where Δx_{av} is the average horizontal distance between shoulders for all rest frames i.e. $f = 1$ and has the value 81.87, and Δx^s is the shoulders horizontal distance for sentence s frame set. The horizontal size normalization becomes $x^* = x_p^f \times \frac{\Delta x_{av}}{\Delta x^s}$ where x^* is the normalized x for both position and size The vertical size normalization is basically the same, except that we calculated the height of the body as the average vertical distance from the face centroids and the two hands centroids of the initial frame (rest frame). We then calculated the normalization factor for each frame as $\frac{\Delta y_{av}}{\Delta y^s}$ where Δy_{av} is the average vertical distance between the face and the hands for all rest frames i.e. $f = 1$ and has the value 114.22, and Δy^s is the shoulders horizontal distance for sentence s frame set. The horizontal size normalization becomes $y^* = y_p^f \times \frac{\Delta y_{av}}{\Delta y^s}$ where y^* is the normalized y for both position and size.

2.3 Learning and recognition

For recognition we used HMM's of different numbers of states, all of which have left to right topology. The feature vector we used has 24 data points. The first 8 points are the hand centroids x and y coordinates for both the left and right camera. The next 8 points are the velocity of the hand centroids in both the x and y directions and in stereo. The last 8 points are the acceleration of the centroids in the x and y directions and in stereo. In each run we used a training set of 3 persons and one set for recognition, doing all the possible combinations within this constraint. We also

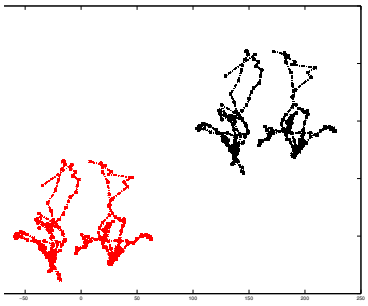

Figure 3: Comparison of size and position of the normalized data (red line) and the non-normalized data (black line) for female 1, sentence 1, recording 1, left camera.

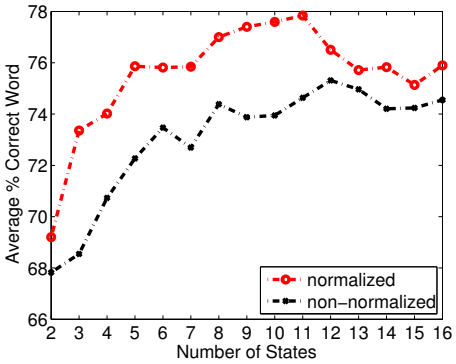

Figure 4: Comparison of average correct word percentage before and after normalization

changed the number of states from 2 to 16 with a unit increment. Figure 4 shows comparisons of different runs in different settings.
As can be seen from figure 4 that there is a persistent increase in the recognition rate for all the settings.

3. Conclusion

We conclude that in the mentioned experimental settings, normalizing data in terms of position and size does increase the recognition rate of the system. This increase is persistent albeit small namely 3.41% in average. In real settings where body sizes vary greatly (e.g With the presence of child signers) the effect of normalization on recognition is expected to be more pronounced. The current work can be improved by devising a better correction algorithm, that utelizes arm and forearm information to achieve better hand position prediction and normalization.

References

[1] H. Yabe et. al. Rwc database –gesture databse–. *IEICE technical report of IE*, 100(179):45–50, 2000.

[2] S. Yanagi, Y. Yagyu, K. tokuda, and T. Kitamura. Hmm-based sign language recognition using hand gesture and hand posture. *Proc. of IEICE General Conference*, page 285, Mar 2003.