PAPER REVIEW: HAND GESTURE RECOGNITION METHODS

Assoc. Prof. Abd Manan Ahmad 1, Dr Abdullah Bade 2, Luqman Al-Hakim Zainal Abidin 3

1 Department of Computer Graphics and Multimedia, Faculty of Computer Science and Information System, Universiti Teknologi Malaysia
81310 Skudai, Johor.

2 Department of Software Engineering, Faculty of Computer Science and Information System, Universiti Teknologi Malaysia
81310 Skudai, Johor.

3 Department of Information Systems, Faculty of Computer Science and Multimedia, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor.

manan@utm.my 1, abade@utm.my 2, luqman@uthm.edu.my 3

ABSTRACT

There are many approaches and algorithms that can be used to recognize and synthesize the hands gesture. Each approach has its own advantages and characteristics. This paper describes the usage of Hidden Markov Models (HMM) and Principal Component Analysis (PCA) in recognizing hands gesture by two different researches. The limitations of each techniques and comparisons between each other will be detailed below.

Keywords: hand gesture recognition, computer vision

1 INTRODUCTION

Sign language is the communication tool for the deaf community around the world. The sign language is a combination of hand gestures, movement sequences and facial expressions.

Machine sign language recognition appeared in the literature since 1990. In its early development, small training sets and isolated signs were used. An image processing system by Tamura and Kawasaki [1] succeeded in recognizing 20 Japanese signs by matching cheremes.

Until now, most work that involves sign language recognition used the data gloves which is worn by the user [2][3]. These systems have been focusing on finger signing, that is the user spells the word with hand signs alphabetically [4]. However, if this kind of systems is to be applied in everyday use is not practical because of too many words to be spelled where actually they can be described as gestures (that represent the whole object or words).

Since that, we can find works that able to translate the hand gesture into meaning without having to spell alphabet by alphabet [5]. Their system suggested that the user does not have to wear the data gloves but instead only solid color gloves to help the hand tracking frame rate and stability. Next the shape, orientation and trajectory information is sent to input to a Hidden Markov Model (HMM) to be recognized as words. They succeeded at 99% accuracy out of 40 training gestures.

From this paragraph onwards, the Hidden Markov Model will be written as HMM, American Sign Language as ASL and principal component analysis as PCA.

Research by Aalborg University [6] in the other hand, uses the principal component analysis (PCA) method in order to recognize 25 gestures of international hand alphabet in real time. The PCA is used to extract the features from the hand gesture’s images. PCA has been widely used in a variety of recognition issues and the documented theories are documented [7][8][9]. By detailing on choosing the
pixels count/numbers, image resolution, training images number and the number of features as parameters in PCA design, the system recognized 99% gestures in online and offline.

Tony Heap and Fernando Samaria [10] were detailing on the usage of Active Shape Models (Smart Snakes) to track hand and a genetic algorithm for initial global image searching. The Smart Snakes is being used in locating feature within a still image by placing the shape (of the feature) closely to it. The contour is attracted to nearby edges in the image and made to move towards the edges to fit the feature. This method however has one disadvantage, which is capturing the way how the image can vary and at the same time disallowing invalid deformation. In short, this drawback can be taken care of by using the Point Distribution Model as in Fig 1 [10].

2 HIDDEN MARKOV MODEL IN VISUAL RECOGNITION OF AMERICAN SIGN LANGUAGE

HMM is a statistical tool for modeling a wide range of time series data. Traditionally, HMM has been used widely in the domain of speech recognition [11]. HMM can be derived from a Markov Model, but the difference is that HMM is much more expressive and represent a better prediction. This paper will be reviewing the usage of HMM in recognizing ASL on the first person view by Thad Starner [11]. The first person view system is chosen because it is more accurate rather than the second person view method.

In specifying a sign, the initial topology for the HMM is determined by estimating the number of different states involved. A five states of HMM is used because it is considered sufficient even for the most complex sign. In order to handle less uncomplicated sign, a skip transitions were specified so that the topology is allowed to imitate HMM with either 3 or 4 states. The author however determined to design a 4 state HMM with 1 skip transition after testing several different topologies as in Fig. 2 below by [11].

2.1 The first person view method: Capturing and Detecting the Hand Region

Using a camera mounted on a hat, the system captured and analyzed color NTSC composite video at 320 by 243 pixel resolutions. The hand tracking is maintained at 10 frames per second, in which is sufficient for human recognition [12].

2.2 The first person view method: Why is it better than the second person view

Table 1 [5] shows that wearable computer system achieved up to 97.8% recognition rate of ASL comparing to 94.1% recognition rate if using the desk based system. The second person view is built by using a stand camera or webcam facing the sign language signer. This method is less accurate than the other because of more occlusion problem with the face and the hand; problem with body rotation, each data set was created and verified by
separate subjects with successively better data recording methods.

Table 1. Comparison of ASL Recognition Accuracy between First and Second Person View Method

<table>
<thead>
<tr>
<th>Desk based system (second person view method)</th>
<th>Wearable computer system (first person view method)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All features</strong></td>
<td><strong>5 word sentences</strong></td>
</tr>
<tr>
<td><strong>Training set</strong></td>
<td><strong>Training set</strong></td>
</tr>
<tr>
<td>94.1 %</td>
<td>99.3 %</td>
</tr>
<tr>
<td><strong>Independent test set</strong></td>
<td><strong>Independent test set</strong></td>
</tr>
<tr>
<td>91.9 %</td>
<td>97.8 %</td>
</tr>
<tr>
<td><strong>Relative features</strong></td>
<td><strong>Part-of-speech</strong></td>
</tr>
<tr>
<td>87.2 %</td>
<td></td>
</tr>
</tbody>
</table>

3 PRINCIPAL COMPONENT ANALYSIS

The basis of PCA is reducing the variables number of a dataset which leads to minimizing the dimension of classification problem. As defined in [13], the PCA is to reduce the dimensionality of a large number (of interrelated variables) in a dataset. According to [14], PCA at same time can retain as much variation as possible in that particular dataset by transforming the principal components to a new set of uncorrelated variables.

A research by [15] proved a relationship in the joint angle measurements as in Fig 4[15]. And therefore, the PCA is applied to reduce the search space while preserving the components with the maximum energy. The usage of PCA in their research resulted in producing a compact 3-dimensional hand model.

3.1 Real Time Recognition of Hand Alphabet Gestures Using PCA : Normalizing and Transforming Hand Image

A PCA is also useful in determining which features are more important in the classification. Although PCA is widely used, many researchers fail to analyze the different parameters usage that has dissimilar outcome in designing the system. A paper by [6] focused on choosing the important parameters involved in designing a PCA based system. They developed a system that is divided into two parts; offline and online part which lead to 99% recognition rate on 1500 images. The offline part is performed to find the transformation matrix and generate a classifier based on a set of training images. Next, the online part uses the transformation matrix with the classifier (from the offline part) to transform and classify new images. The offline part will recognized all the incoming images but when the signer spells a certain word, not all images should be considered as a sign. Therefore the online part works in creating knowledge of the dynamic content in the images series.

As a result of this attribute, the system is very pixel oriented hence make it sensitive towards the position, the orientations and the scale of the hand image. In order to overcome this sensitivity quality, every image is being normalized so that the hand image is located and transformed as in Fig. 5 [6].

3.2 Real Time Recognition of Hand Alphabet Gestures Using PCA : Choosing the Free Parameters

The free parameters chosen in this system is image resolution, the number of features and the number of training images. For the offline processing, the only important parameter is the size of the training data set. For the online processing, the image resolution and the number of features are important. To make the recognition rate achieved better recognition rate, all three parameters are critical. The image resolution plays an important part in recognizing an image. Despite using higher
resolution images increase the recognition rate and causes longer processing time, an image of 32 by 32 pixels is sufficient and therefore is used to recognize the hand image.

Other parameters being considered is the number of training images. It is expected and proven by experimental result that more training image increase the recognition rate but how many is more? A test was conducted and 40 training images are enough to achieve 98.4% of recognition rate as in Fig 6 [6] comparing to 95.7% recognition rate by using 20 training images.

As discussed in 3.1, the offline part of the system will process all incoming images to be recognized and the online part will create knowledge of the dynamic content. In spelling the word ‘pca’ in 2 seconds, 30 frames of images are captured. Other than the gesture of the alphabet P, C and A there are also other transition images captured. This captured transition images is not needed and thus causing confusion in producing the meaningful gestures meant. As in Table 2 [6], the offline part will be translating the gestures as ‘i, p, p, o, p, p, p, p, p, p, p, p, n, i, n, c, c, c, c, c, c, p, p, a, a, a, a instead’ of ‘pca’.

Table 2 : The recognized gestures during spelling PCA

<table>
<thead>
<tr>
<th>frame</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
</table>

Figure 6. The recognition rate over the number of selected features. The symbol o, *, = and x is the number of training images used.

The number of features to be used is the most important parameters for the recognition rate graphs. The more selected features being used, the image became clearer. However, at a certain number of features selected the recognition rate stops increasing. There is no distinct image being retrieved after using more than 16 features to re-map the image. Fig. 7 [6] shows that from the input image, the shape of the hand is recognizable at 15 and 20 features selected rather than select only 1, 2 or 10 features.

In order to solve this temporal segmentation problem, 2 motion cues are introduced in the feature space. The first cue will be assigned a value of 0 if there is a motion of the hand’s centre of mass (COM) involved and the second cues is assigned with the value of 1 if there is a motion of changing gestures detected in the image series. The third cue is introduced so that the system can spell a character twice if intended by the signer, such as the word ‘ally’. It is the result of inverted ANDing of the first and second cues. Finally, all three cues are ANDED together to recognize the gesture in a frame if the result is 1 as in Table 3 [6].

Table 3 : The word PCA successfully recognized

<table>
<thead>
<tr>
<th>frame</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>cue 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>cue 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>cue 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>recognized</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>frame</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>30</td>
<td>31</td>
<td>32</td>
</tr>
<tr>
<td>cue 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>cue 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>cue 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>recognized</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

4 CONCLUSION AND DISCUSSION

4.1 DISCUSSION

As discussed in the previous chapter, the usage of HMM and PCA helped in recognizing the hand gesture. However, in designing a robust and effective system to recognize hand gesture, which technique has the highest recognition rate? The discussion of both models will be detailed in this section.
4.2 Comparison of Hidden Markov Model and Principal Component Analysis in Recognizing Hand Gesture

Though the wearable sign language recognizer by Thad Starner [5] able to achieve 99.7% recognition rate, there is a few thing that needed to be looked upon further to make is unconstrained. As the need for the system to be used widely by many signers, it must be designed to be able to address how many fingers are visible along the hand contour and whether the palm orientation is facing up or down.

Another character of the system is that it is able to solve the finger spelling properties and also face recognition. There are words in sign language that must be combined together with the face mimes other than the hand gestures. In a larger training set and context modeling, it is expected the system to have higher error rates and thus makes it user dependent. In order to fix this, the hand position towards the body should be measured [17] and to integrate the head tracking with the facial gestures in the feature set.

The PCA based system in [6] is sensitive towards the changes the hand gestures pose. As PCA is a data driven method, this sensitivity attribute is solved by normalizing the hands region but it is still responsive to orthogonal rotations of the camera axis and the momentous change in signing a gesture.

4.3 Conclusion and Future Work

This paper is to review which technique between PCA and HMM to be applied in recognizing hands gesture and sign language.

The application of HMM by Thad Starner in his previous research [5][11][17] proved that HMM is a powerful tool in recognizing hand gesture and sign language up to 99.8% recognition rate.

In the other hand, the usage of PCA by fellow researchers [6][16] also proved that it is also able to recognize hand gesture and sign language.

For future work, it is advisable to combine the HMM method together with the PCA in recognizing hand gesture or sign language because both are a very powerful tools in this study domain.

In conclusion, a robust hand gesture or sign language recognizer system should be handled in real-time, user independent, able to be used by a large number of signer and fast training method.

4.4 Acknowledgement

This research is supported by the Ministry of Science, Technology and Information (MOSTI) Malaysia under the VOT number 01-01-06-SF0507, “The Development of Hand Gesture Interpreter for Standard Malay Text”.

REFERENCES


