Sign Language Recognition based on Key Frame

Shengwei Zhang*, Zhaosong Zhu and Zuojin Hu
Nanjing Normal University of Special Education, Nanjing, China

*Corresponding author email address: henzy@qq.com

Abstract. This paper presents a new real-time recognition system as for Chinese sign language. In order to decrease massive unnecessary data, including the body of model, the differential method is used to extract the arms of model. The sign language videos are stored in the different groups, where the number of key frames is same, to decrease matching process. Hu invariant moments models are implying to extract the image’s feature and obtain the key frames.

1. Introduction

Sign Language Recognition (SLR) aims to develop algorithms and methods to correctly identify a sequence of produced signs and to understand their meaning. [1] Sign consists of three main parts: Manual features involving gestures made with the hands (employing hand shape and motion to convey meaning), Non-manual features such as facial expressions or body posture, which can both form part of a sign or modify the meaning of a manual sign, and Finger spelling, where words are spelt out gesturally in the local verbal language, which in China is Chinese phonetic alphabet. Sign language is as complex as any spoken language, each sign language has many thousands of signs, each differing from the next by minor changes in hand shape, motion, position, non-manual features or context. While automatic speech recognition has now advanced to the point of being commercially available, automatic Sign Language Recognition is still in its infancy. Because of wide country and multiple people, there are many differential forms for only one sign language in Chinese sign language.

Charaphayan and Marble [2] Attempts to automatically recognize sign language began to appear at the end of 80’s. They investigated a way using image processing to understand American Sign Language (ASL). This system can recognize correctly 27 of the 31 ASL symbols. T. Starner achieved a correct rate of 91.3% for 40 signs based on the image. By imposing a strict grammar on this system, the accuracy rates in excess of 99% were possible with real-time performance. Fels and Hinton developed a system using a VPL Data Glove Mark II with a Polhemus tracker as input devices. Neural network was employed for classifying hand gestures. Y. Nam and K.Y. Wohn used three–dimensional data as input to HMMs for continuous recognition of a very small set of gestures. They introduced the concept of movement primes, which make up sequences of more complex movements. R. H. Liang and M. Ouyang used HMM for continuous recognition of Tainwan Sign language with a vocabulary between 71 and 250 signs by using Data glove as input devices. HMMs were also adopted by Kisti Grobel and Marcell Assan to recognize isolated signs collected from video recordings of signers wearing colored gloves, and 91.3% accuracy out of a 262-sign vocabulary was reported. C. Vogler and D. Metaxas [3] used HMMs for continuous ASL recognition with a vocabulary of 53 signs and a completely unconstrained sentence structure. C. Vogler and D. Metaxas described an approach to continuous, whole-sentence ASL recognition that used phonemes instead of whole signs as the basic units. They experimented with 22 words and achieved similar recognition rates with phoneme-based and word-based approaches. Wen
GAO [4] proposed a Chinese Sign language recognition system with a vocabulary of 1064 signs. The recognition accuracy is about 93.2%.

In this paper, we use method of difference and filter to extract the figure of arms, define the key frame according to Hu invariant moments, and classify sign in term of the optimal path searching techniques.

2. Extract the figure of the arms

Key frame, in this paper, is defined to describe the change of hand shape in the movement of sign language. Analyzing larges of sign language videos, we find that there are approximately 1-4 seconds in each videos. If we set photographing frequency of Kinect as 30 frames per second, we can obtain about 30-120 images for each sign language. There are many redundancy in these images only to present transition of the hand shapes, which do not present the practical significance. So we extract the key frame to cut down the unnecessary redundancy and confirm the video as one kind of sign languages in the next step.

Through sampling and quantization, the result of a video is real matrix. For convenience, we adopt an integral array to describe a digital image as follow:

\[
\begin{pmatrix}
    f(1,1) & f(1,2) & \cdots & f(1,N) \\
    f(2,1) & f(2,2) & \cdots & f(2,N) \\
    \vdots & \vdots & \ddots & \vdots \\
    f(M,1) & f(M,2) & \cdots & f(M,N)
\end{pmatrix}
\]  

In our sign language corpus, we find that the body of sign language models almost remain motionless, or move slightly. We can use difference of the 2 successive image of sign language video to cut off major data of model’s body, and adopt filter to obtain the figure of the arms of the model.

3. Shape Feature Description based on Hu’s Moments

In order to retrieve image accurately and quickly, effective feature description method is necessary. For the reason that the methods based on feature points cannot describe the shape information of object. Hu invariant moments of edge are used to describe the shape feature in the next step. Hu invariant moments, also called geometric invariant moments are widely used in image processing because the robustness to image translation, scale and rotation transformation. In this paper, Hu invariant moments are calculated to describe the shape feature based on the differenced sign language images.

Firstly, for image \( f(x, y) \) with size \( M \times N \), the \( p+q \) order moments are defined as

\[
m_{pq} = \sum_{m=1}^{M} \sum_{n=1}^{N} x^p y^q f(x, y)
\]  

Then the central moments are defined as

\[
\mu_{pq} = \sum_{m=1}^{M} \sum_{n=1}^{N} (x - \bar{x})^p (y - \bar{y})^q f(x, y)
\]  

Where \( \bar{x} = m_{10}/m_{00}, \bar{y} = m_{01}/m_{00} \).

In order to normalize the central moment \( \eta_{pq} \), each \( \eta_{pq} \) is divided by \( \mu_{00}^p \). The normalized central moment is shown as

\[
\eta_{pq} = \frac{\mu_{pq}}{(\mu_{00})^p}
\]  

Where \( \rho = \frac{p+q}{2} + 1 \).

Finally, the 7 order Hu invariant moments are defined as follows:
The sign image can be transformed to the binary image according to calculate its Hu invariant moments. As for all of images of the each sign language video, we can obtain their key frames as follows:

(1) Suppose that $I = \{I_1, I_2, ..., I_n\}$ presents the differed image of the sign language video, we use Hu invariant moments to process them and get the normalized result as $H_{ui}$, where $H_{ui} = \{h_{ui1}, h_{ui2}, h_{ui3}, h_{ui4}, h_{ui5}, h_{ui6}, h_{ui7}\}$. Based on the designated frame $I_q$, the Euclidean distance between $I_q$ and $I_1~I_n$ will be calculated as $D_q = \{d_{q1}, d_{q2}, ..., d_{qn}\}$. In this paper, we based on $I_1$, sort all the elements of $D_q$, form the sorted sequence, namely, $d_{q1} \leq d_{q2}$, and record the original serial number as the array $P = \{P_1, P_2, ..., P_n\}$.

(2) The sorted sequence $D_q$ will be section into 2 classes, $T_1 = \{d_1, d_2, ..., d_{q}\}$ and $T_2 = \{d_{q+1}, d_{q+2}, ..., d_n\}$, then calculate the distance $E(u)$ between this 2 classes and the $E(u+1)$ that select the element $d_{q+1}$ from $T_2$ to add into $T_1$. Comparing the $E(u)$ and $E(u+1)$, if $E(u+1) > E(u)$, we will repeat the last step and add the element from $T_2$ into $T_1$, until $E(u+x) < E(u+x-1)$. The $d_{q+x}$ will be record to describe the frame serial number of images, where the corresponding original image $P_i$ is defined as key frame.

(3) $P_1$ is described as the first key frame of total sign language video. The rest, $P_1~P_n$, will be regarded as a section. In term of the last two steps (1) and (2), the next key frame will also be confirmed, and so on. If there is only one element in $T_2$, its corresponding original image will be regarded as the key frame, and the extracting process is over.

4. Sign language video match

Through key frame extracting, all of the sign language videos in our experiment are classified into different groups according to the number of the video’s key frames, and each group can have a different number of sign videos, which store in their key frames with sign label as for reference patterns. When the experimenter demonstrate his/her sign language, its video will be processed as the previous steps 2 and 3 to obtain its key frame and also the number of them. The matching works as follow:

Let $r(i), i = 1, 2, ..., I$ and $t(j), j = 1, 2, ..., J$ be the respective feature vector sequences for a specific pair of reference and test patterns, where in general $I \neq J$, so that measures based on optimal path searching techniques will be adopted to develop an appropriate distance between the two sequence. A two-dimensional grid is formed with the elements of two sequences as points on the respective axes, namely, the reference array at the abscissa (i-axis) and the test one at the ordinate (j-axis). Each node $(i, j)$ of the grid is associated with a cost, which is an respective defined function $d(i, j)$ measuring the ‘distance’ between the respective elements of the arrays, $t(j)$ and $r(i)$. A path through the grid from the initial node to the final one is an ordered set nodes and the overall cost $D$ is defined as

\[
\begin{align*}
M_1 &= n_{i0} + n_{i2} \\
M_2 &= (n_{i0} + n_{i2})^2 + 4n_{i1}^2 \\
M_3 &= (n_{i0} - 3n_{i2})^2 + (3n_{i2} - n_{i0}) \left( n_{i0} + n_{i2} \right) - 3(n_{i2} + n_{i0}) \\
M_4 &= (n_{i0} - n_{i2})^2 + (n_{i2} - n_{i0}) \left( n_{i0} + n_{i2} \right) - 3(n_{i2} + n_{i0}) \\
M_5 &= (n_{i0} - 3n_{i2})^2 (n_{i0} + n_{i2}) \left( n_{i0} + n_{i2} \right) - 3(n_{i2} + n_{i0}) \left( n_{i0} + n_{i2} \right) \\
M_6 &= (n_{i0} - n_{i2})^2 (n_{i0} + n_{i2}) \left( n_{i0} + n_{i2} \right) - 3(n_{i2} + n_{i0}) \left( n_{i0} + n_{i2} \right) \\
M_7 &= (n_{i0} - 3n_{i2})^2 (n_{i0} + n_{i2}) \left( n_{i0} + n_{i2} \right) - 3(n_{i2} + n_{i0}) \left( n_{i0} + n_{i2} \right) - (n_{i0} - 3n_{i2})^2 (n_{i0} + n_{i2}) \left( n_{i0} + n_{i2} \right) - 3(n_{i2} + n_{i0}) \left( n_{i0} + n_{i2} \right) \\
\end{align*}
\]
\[ D = \sum_{k=0}^{K-1} d(i_k, j_k) \] (6)

Where \( K \) is the number of nodes along the path.

We will calculate the overall cost in the group with the same numbers of key frame and regard the label with least cost is the sign language that the test video shows.

5. Conclusion
The paper introduces a whole set of techniques for achieving real time sign language recognition, with the following important process:

- Gesture recognition using both arms is implemented in the differential of successive frame or image to filter the body of the sign language model. This can be further enhanced by increasing the number of gestures in the reference data, at the cost of slower computation.

- Hu invariant moment’s models are implying that the recognition works regardless of the geometric feature of the person performing the sign.

Acknowledgements
This research was supported by Natural Science Foundation of Colleges in Jiangsu Province (No. 16KJB520026), (No. 16KJB520029) and (No. 17KJB520019).

References
[1] Cooper, Helen, B. Holt, and R. Bowden. Sign Language Recognition. Visual Analysis of Humans. Springer London, 2011: 539 - 562.
[2] Charayaphan C, Marble A E. Image processing system for interpreting motion in American Sign Language [J]. Journal of Biomedical Engineering, 1992, 14 (5): 419.
[3] Vogler C, Metaxas D. ASL Recognition Based on a Coupling Between HMMs and 3D Motion Analysis [C]// International Conference on Computer Vision. IEEE, 1998: 363 - 369.
[4] Chunli Wang, Wen GAO, Shiguang Shan. An approach based on phonemes to large vocabulary Chinese sign language recognition [M]. 2002.
[5] Guo, Gongde, et al. KNN Model-Based Approach in Classification. On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. Springer Berlin Heidelberg, 2003: 986 - 996.
[6] Min-Ling Zhang, Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40 (7): 2038 - 2048.
[7] Sergios Theodoridis, Konstantinos Koutroumbas. Pattern Recognition. China Machine Press, 2010.
[8] Han J, Shao L, Xu D, et al. Enhanced Computer Vision With Microsoft Kinect Sensor: A Review [J]. IEEE Transactions on Cybernetics, 2013, 43 (5): 1318 - 1334.
[9] S. Milani and G. Calvagno, Joint denoising and interpolation of depth maps for MS Kinect sensors, in Proc. ICASSP, 2012, pp. 797 – 800.
[10] Kumar A, Thankachan K, Dominic M M. Sign language recognition [C]// International Conference on Recent Advances in Information Technology. IEEE, 2016: 422 - 428.