# A Sign Language Recognition Based on Tensor

Su-Jing Wang, De-Cai Zhang, Cheng-Cheng Jia, Na Zhang, Chun-Guang Zhou, Li-Biao Zhang*
College of Computer Science and Technology, Jilin University, Changchun 130012, P. R. China

***Abstract*—The hand gesture recognition plays a key role in many appealing applications. The sign language recognition is one of the important applications of hand gesture recognition. The existing methods on sign language recognition are limited to certain view. In this paper, we use tensor subspace analysis to model a multi-view hand gesture to recognize 26 manual alphabetical letters. In our experiment, each hand gesture is captured from 5 different views. Two experiments are conducted on gray-scale images and binary images, respectively. The results show the proposed method has a good performance on multi-view.**

## I. Introduction

In pattern recognition and computer vision, hand gesture recognition plays a key role in many appealing applications. For instance, Some novel application environments include virtual and augmented reality, interaction with large displays, enhanced visualization of large data collections to name a few. In such cases, the keyboard and mouse pair can show its limits. From this perspective, the use of hand gestures for Human-Computer Interaction can help people to communicate with computer-based systems in a more intuitive way. Furthermore, hand gesture recognition yet can be applied to the interpretation and learning of sign language (SL). SL is the dominant communication medium for the deaf community. SL recognition (SLR) which aims at automatically transcribing signs into text or speech by means of computer has gained growing attention over the past 15 years.

Existing methods on SLR can be divided into 2 categories by ways of data acquisition: the SLR based on data glove and the SLR based on vision. The former can sample more accurate data than the later. But the data glove is expensive, bulky, flimsy and inconvenient. Comparing with the SLR based on data glove, the SLR based on vision is convenient and naturally. So, it is significant to study on the SLR based on vision.

However, the most of existing SLR based on vision are limited to certain view. Starner and Pentland[1] used Hidden Markov Model to study on SLR. Each eigenvector in their model consists of 8 components, such as x, y coordinates of hands, the orientationes of principal axes and the eccentricities of circumscribed ellipses of hands. These components depend on a view. Bowden *et al.*[2] proposed a group of linguistic features to recognize SL without the positions of cameras, to some extent. But the ways to get this linguistic features are relative to the positions of cameras. The above researches on SLR are limited to a certain view. This leads to using inconveniently the SLR based on vision.

The restriction of viewing from a particular posture means the hand gesture users can only use hand languages in confined orientations within confined space, which is clearly inconvenient for them. Therefore, it is necessary to explore the feasibility of orientation-free hand gesture recognitions.

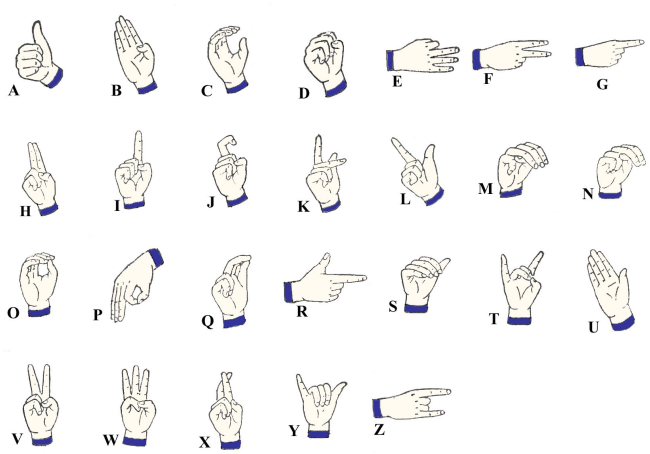



Fig. 1. manual alphabet

Using tensor subspace analysis, we model a multi-view hand gesture to recognize 26 manual alphabet, which are showed in Fig.1. In a multilinear analysis framework, a hand gesture tensor consisting of hand gestures for all the hand gestures to be recognized in multiple views can be factorized into a Tucker tensor production of the core tensor and a hand gesture basis matrix and a view basis matrix using high-order singular value decomposition[3].

## II. Tensor Fundamentals

A tensor is a multidimensional array. More formally, an *N*th-order tensor is an element of the tensor product of *N* vector spaces, each of which has its own coordinate system. In this paper, lowercase italic letters (*a, b, ...*) denote scalars, bold lowercase letters (**a, b, ...**) denote vectors, bold uppercase letters (**A, B, ...**) denote matrices, and calligraphic uppercase letters ($\mathcal{A}$, $\mathcal{B}$, ...) denote tensors. The formal definition is given below:

*Definition 2.1:* The *order* of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is $N$. The *mode-n vectors (fibers)* of $\mathcal{A}$ are the $I_n$-dimensional vectors obtained from $\mathcal{A}$ by fixing every index but index $i_n$.

The mode-n vectors of $\mathcal{A}$ is given in Fig.2 and the flattened matrix of $\mathcal{A}$ is given in Fig.3.

*Definition 2.2:* The *mode-n product* of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ by a matrix $\mathbf{U} \in \mathbb{R}^{J_n \times I_n}$, denoted by $\mathcal{A} \times_n \mathbf{U}$, is an $(I_1 \times I_2 \times \ldots \times I_{n-1} \times J_n \times I_{n+1} \times \ldots \times I_N)$-tensor of

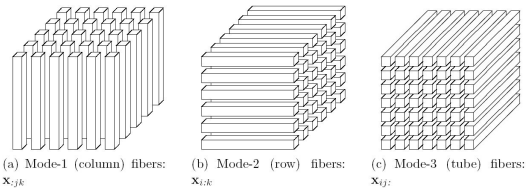



Fig. 2. Fibers of A 3rd-Order Tensor

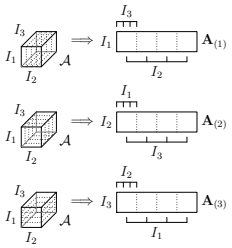



Fig. 3. Flattening of the $(I_1 \times I_2 \times I_3)$-tensor $\mathcal{A}$ to the $(I_1 \times I_2 I_3)$-matrix $\mathbf{A}_{(1)}$, the $(I_2 \times I_3 I_1)$-matrix $\mathbf{A}_{(2)}$, and the $(I_3 \times I_1 I_2)$-matrix $\mathbf{A}_{(3)}$ $(I_1 = I_2 = I_3 = 4)$.

which the entries are given by

$$(\mathcal{A} \times_n \mathbf{U})_{i_1 i_2 \ldots i_{n-1} j_n i_{n+1} \ldots i_N} \stackrel{\text{def}}{=} \sum_{i_n} a_{i_1 i_2 \ldots i_{n-1} i_n i_{n+1} \ldots i_N} u_{j_n i_n}. \tag{1}$$

This mode-n product of tensor and matrix can be expressed in terms of unfolding matrices for ease of usage.

$$(\mathcal{A} \times_n \mathbf{U})_{(n)} = \mathbf{U} \cdot \mathbf{A}_{(n)} \tag{2}$$

Given the tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ and the matrices $\mathbf{U} \in \mathbb{R}^{J_n \times I_n}$, $\mathbf{V} \in \mathbb{R}^{J_m \times I_m}$, one has

$$(\mathcal{A} \times_n \mathbf{U}) \times_m \mathbf{V} = (\mathcal{A} \times_m \mathbf{V}) \times_n \mathbf{U} = \mathcal{A} \times_n \mathbf{U} \times_m \mathbf{V} \tag{3}$$

Similarly, *N*-mode SVD is a generalization of the SVD for higher order matrices[3]. If $\mathcal{D}$ is an *n* order tensor and $\mathcal{D} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, the application of *n*-mode SVD orthogonalizs "n" associated vector spaces of $\mathcal{D}$ and decomposes the tensor as

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \ldots \times_n \mathbf{U}_n \ldots \times_N \mathbf{U}_N \tag{4}$$

where $\mathbf{U}_n$, $\forall n \in \{1, 2, \ldots, N\}$, is an orthonormal matrix and contains the ordered principal components for the *n*th mode. $\mathcal{Z}$ is called the *core tensor*. The decomposition algorithm is as follows:

1) For $n = 1, \ldots, N$, compute matrix $\mathbf{U}_n$ in (4) by computing the SVD of the flattened matrix $\mathbf{D}_{(n)}$ and setting $\mathbf{U}_n$ to be the left matrix of the SVD.
2) Solve for the core tensor as follows:

$$\mathcal{Z} = \mathcal{D} \times_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2^T \ldots \times_n \mathbf{U}_n^T \ldots \times_N \mathbf{U}_N^T \tag{5}$$

## III. Proposed approaches

### A. Hand region detection

The first step of a hand recognition process is the detection of the hand region. In the proposed method, this is achieved through color segmentation, *i.e.* classification of the pixels of the input image into skin color and non-skin color clusters. The technique is based on color information, because color is a highly robust feature.

The YCbCr color space was developed as part of ITU-R BT.601. It is used as a part of the color image pipeline in video and digital photography systems. Y is the luma component, Cb and Cr are the blue-difference and red-difference chroma components.

Although RGB color space is considered adequate for most consumer applications, due to high relativity with luminance, it is not suitable for the segmentation in the color space. The Experiment in the paper[4] showed that Cb and Cr values are narrowly and consistently distributed in the maps of the chrominance components of skin color, and the YCbCr color space is not sensitive to the luminance. So we will use YCbCr color space in the detection of skin rather than the RGB color space.

Digital YCbCr is derived from digital RGB (8 bits per sample) according to the following equations:

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65.481 & 128.553 & 24.966 \\ -37.797 & -74.203 & 112 \\ 112 & -93.786 & -18.214 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{6}$$

In this paper we choose to use the human skin color model, which was proposed by Chen *et al.*[5], for skin detection. The model they proposed uses BP networks and is robust for various illuminations.

### B. modelling SLR based using tensor

Given the multi-view manual alphabet images rasterized as a tensor $\mathcal{D} \in \mathbb{R}^{I_l \times I_v \times I_{pix}}$, where $I_l$, $I_v$ and $I_{pix}$ denote the number of letters, views and pixels, respectively. HOSVD is applied to factorize letter and view information *etc*. as Eq.(7),

$$\mathcal{D} = \mathcal{Z} \times_1 \mathbf{U}_l \times_2 \mathbf{U}_v \times_3 \mathbf{U}_{pix} \tag{7}$$

where the *core tensor* $\mathcal{Z} \in \mathbb{R}^{I'_l \times I'_v \times I'_{pix}}$ governs the interaction among the factors represented in the 3 mode matrices. The mode matrix $\mathbf{U}_l \in \mathbb{R}^{I_l \times I'_l}$ and $\mathbf{U}_v \in \mathbb{R}^{I_v \times I'_v}$ span the parameters space of various letters and views, respectively. The mode matrix $\mathbf{U}_{pix} \in \mathbb{R}^{I_{pix} \times I'_{pix}}$ constitute the space of *eigenhands*. The $l$-th row of $\mathbf{U}_l$, denoted by $\mathbf{u}_l^{(l)}$, is the coefficient vector of identity $l$. The $v$-th row of $\mathbf{U}_v$, denoted by $\mathbf{u}_v^{(v)}$, is the coefficient vector of view $v$.

From Eq.(7), a training letter image $\mathcal{D}^{(l,v)}$ for the letter $l$ and the view $v$ is

$$\mathcal{D}^{(l,v)} = \mathcal{Z} \times_1 \mathbf{u}_l^{(i)} \times_2 \mathbf{u}_v^{(v)} \times_3 \mathbf{U}_{pix} \tag{8}$$

Here, $\mathcal{D}^{(l,v)}$ is a $1 \times 1 \times I_{pix}$ tensor. We reconstruct hand gesture images by fixing $v$, varying $l$ and selecting the first 1st, 5th, 10th, 50th, 100th, 200th eigenvectors form $\mathbf{U}_{pix}$, which are illustrated in Fig.4. In the same way, various views with the increasing number of eigenvectors are showed in Fig.5.

According to Eq.(3), Eq.(7) can be transformed as follows:

$$\begin{aligned} \mathcal{D} &= (\mathcal{Z} \times_2 \mathbf{U}_v \times_3 \mathbf{U}_{pix}) \times_1 \mathbf{U}_l \\ &= \mathcal{B} \times_1 \mathbf{U}_l \end{aligned} \tag{9}$$

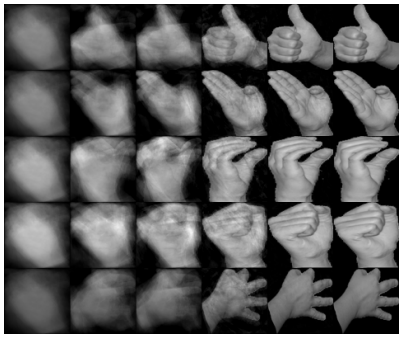


Fig. 4. various letters with the increasing number of eigenvectors

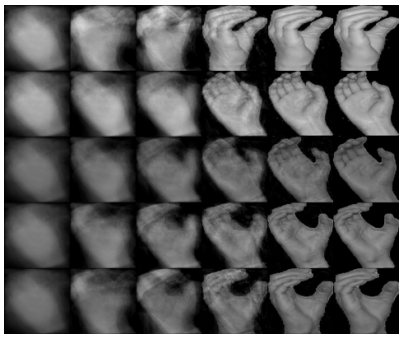


Fig. 5. various views with the increasing number of eigenvectors

### C. recognizing SL based on tensor

The recognizing SL method is extend from MPCA-LV[6]. For a test hand gesture image $\mathcal{D}_{test}$, we have,

$$\mathcal{D}_{test} = \mathcal{Z} \times_1 \mathbf{u}_l \times_2 \mathbf{u}_v \times_3 \mathbf{U}_{pix} \tag{10}$$

where $\mathbf{u}_l$ and $\mathbf{u}_v$ are letter-space and view-space, respectively. For a test hand gesture image $\mathcal{D}_{test}$, we need to calculate $\mathbf{u}_l$ and $\mathbf{u}_v$. This can be formulated as,

$$\arg\min_{\mathbf{u}_l,\mathbf{u}_v} \|\mathcal{D}_{test} - \mathcal{Z} \times_1 \mathbf{u}_l \times_2 \mathbf{u}_v \times_3 \mathbf{U}_{pix}\| \tag{11}$$

Although we only need $\mathbf{u}_l$, the $\mathbf{u}_v$ still need to be calculated. A fixed set $\{\mathbf{u}_v\}$ is restructured, its size is $I_v$. Let $\mathbf{u}_v^{k_v} \in \{\mathbf{u}_v\}$, Eq.(11) can be rewritten as,

$$\arg\min_{\mathbf{u}_l} \|\mathcal{D}_{test} - \mathcal{Z} \times_1 \mathbf{u}_l \times_2 \mathbf{u}_v^{k_v} \times_3 \mathbf{U}_{pix}\| \tag{12}$$

According to *mode-n flattened matrix*, the above equation can be rewritten as,

$$\arg\min_{\mathbf{u}_l} \|\mathcal{D}_{test} - \mathbf{u}_l \times (\mathcal{Z} \times_2 \mathbf{u}_v^{k_v} \times_3 \mathbf{U}_{pix})_{(l)}\| \tag{13}$$

the above equation can also be rewritten as,

$$\mathbf{u}_l = \mathcal{D}_{test} \times (\mathcal{Z} \times_2 \mathbf{u}_v^{k_v} \times_3 \mathbf{U}_{pix})^+_{(l)} \tag{14}$$

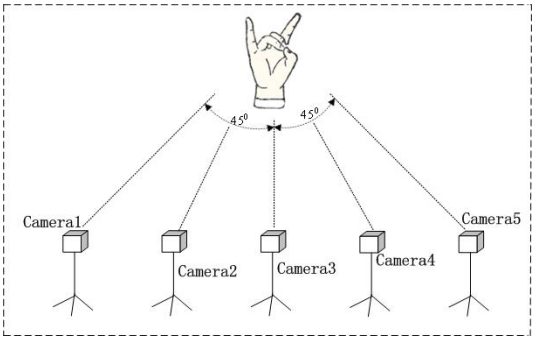



Fig. 6. the positions of 5 cameras

where the superscript + implies *Moore-Penrose pseudoinverse*. For each possible $\mathbf{u}_v^{k_v} \in \{\mathbf{u}_v\}$,

$$\begin{aligned} \mathbf{u}_l^{k_v} &= \mathcal{D}_{test} \times (\mathcal{Z} \times_2 \mathbf{u}_v^{k_v} \times_3 \mathbf{U}_{pix})^+_{(l)} \\ k_v &= 1, \ldots, I_v. \end{aligned} \tag{15}$$

Later, using the cosine distance between $\mathbf{u}_{i_p}^{k_v}$ and $\mathbf{u}_l^{(p)}$ as measurement, the best matching hand gesture is decided by,

$$\arg\max_{l,v} \frac{\langle \mathbf{u}_{i_p}^{k_v}, \mathbf{u}_l^{(p)} \rangle}{\|\mathbf{u}_{i_p}^{k_v}\| \|\mathbf{u}_l^{(p)}\|} \tag{16}$$

## IV. EXPERIMENT

In our experiment, there are 26 hand gestures to express 26 letters. All the hand gesture images are captured from 5 different views. These five cameras are mounted at approximately 2/3 body height with looking directions parallel to the ground plane. This camera setting is illustrated in Fig.6.

There are 130 hand gestures from 5 views are cropped by the minimum exterior rectangle of hand region and then resized to $80 \times 80$ pixels. For different views, the 5-fold cross validation is used to evaluate the performance of this models. In the fold, the i-th(i=1,2,3,4,5) view's images are selected as the testing set, and the remaining are used as the training set. The final result is the sum over the 5 folds. The aim is to analyze the performance of recognizing a hand gesture image from different views. All images are transformed to gray-scale images and binary images, respectively. Fig.7 shows RGB images, gray-scale images and binary images.

For HOSVD and other tensor operations, we used the tensor toolbox developed by Bader and Kolda in MATLAB$^{TM}$[7]. All experiments are conducted on a 2.66 GHz Intel PC with 16 GB main memory, with the Microsoft Windows XP 64 bits as OS.

Two experiments are conducted on gray-scale images and binary images, respectively. The recognition rates for every view are shown in Table.I. From Table.I we can draw the conclusions bellow: (1) The best performance achieves 100% on testing a hand gesture from View 3. The reason is that the person in this viewpoint could see the hand gestures clearly and the images taken in this viewpoint are the most accurate.So the images taken in the other viewpoint may cause bad performance. (2)The performance is better when we

(a) Cropped RGB images

(b) Resized gray-scale images

(c) Resized binary images

Fig. 7. a hand gesture from 5 views on RGB, gray-scale and binary

TABLE I
SL RECOGNITION RATE FOR EVERY VIEW

| Viewpoint | Grey-scale Image | Binary Image |
|---|---|---|
| View 1 | 0.769 | 0.692 |
| View 2 | 0.731 | 0.808 |
| View 3 | 1 | 0.923 |
| View 4 | 0.923 | 0.923 |
| View 5 | 0.923 | 0.885 |
| Mean | 0.869 | 0.846 |

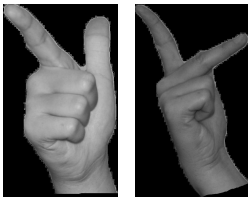
choose the grey-scale images as the test samples rather than the binary images. Obviously, the grey-scale image contains more information than the binary image, such as texture, contour and so on.

TABLE II
SL RECOGNITION RATE FOR EVERY LETTER

| Letter | Grey-scale Image | Binary Image |
|---|---|---|
| A J L P | 4 | 5 |
| B E F G I Q R S U V Y Z | 5 | 5 |
| C H O | 4 | 3 |
| D | 5 | 4 |
| K | 1 | 1 |
| M T Z | 4 | 4 |
| N | 4 | 2 |
| W | 3 | 2 |

As we can see from Table.II, the number of correct recognizing letter K is 1 in times recognition folds. We find that the recognition of letter "L" may cause confusion with Letter "K" when we use grey-scale images as the test samples. Some images of the two letters are illustrated in Fig.8.

Considering the shape, texture and features of these images, the two letters are very much the same. So, we may select more features to enhance the performance of our method in the future research.



(a) Letter *L* (b) Letter *K*

Fig. 8. Letters *L* and *K*

## V. CONCLUSION

In this paper, we use multilinear analysis to model a multi-view sign languages recognition, in order to solve the problem of being limited to a certain view in the existing SLR. The results show the proposed method has a good performance on multi-view. Our future researches will focus on modeling a SLR using mainfold and tensor.

## ACKNOWLEDGMENT

This paper is supported by (1) the National Natural Science Foundation of China under Grant No. 60873146609730926090309 7, (2) National High Technology Research and Development Program of China under Grant No. 2007AAO4Z114, (3) Project of Science and Technology Innovation Platform of Computing and Software Science (985 Engineering), (4) the Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China.(5) National High Technology Research and Development Program of China under Grant No. 2009AA02Z307 (863), (6)the Key Laboratory for New Technology of Biology Recognition of Jilin Province No. 20082209, (7)the Third Phase Construction Project of 211 Engineering of Jilin University. (8)the Science and Technology Development Planning Project of Jilin Province (No.20080168).

## REFERENCES

[1] T. Starner. and A. Pentland., "Visual recognition of american sign language using hidden markov models," in *Proceedings of t he International Conference on Automatic Face and Gesture Recognition*, 1995, pp. 189–194.

[2] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady, "A linguistic feature vector for the visual interpretation of sign language," pp. 390–401, 2004. [Online]. Available: http://www.springerlink.com/content/lnalp43l2r6wcddc

[3] L. De Lathauwer, B. De Moor, and J. Vandewalle, "On the best rank-1 and rank-(r1,r2,...,r-n) approximation of higher-order tensors," *Siam Journal On Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1324–1342, 2000.

[4] E. Stergiopoulou and N. Papamarkos, "Hand gesture recognition using a neural network shape fitting technique," *Engineering Applications of Artificial Intelligence*, vol. 22, no. 8, pp. 1141–1158, Dec. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/B6V2M-4W99NKF-1/2/945564b25eba8cab42ff3fb875683e1b

[5] C. Zhen, W. Su-jing, Z. Chun-guang, and C. Huan-huan, "Human skin color model using bp networks," *Computer Engineering and Applications*, vol. 44, no. 14, pp. 166–168, 2008.

[6] S. Rana, W. Liu, M. Lazarescu, and S. Venkatesh, "A unified tensor framework for face recognition," *Pattern Recognition*, vol. 42, no. 11, pp. 2850–2862, 2009.

[7] B. Bader and T.G.Kolda, "Tensor toolbox version 2.3, copyright 2009, sandia national laboratories, http://csmr.ca.sandia.gov/ tgkolda/TensorToolbox/."