

# A Verification Method for Viewpoint Invariant Sign Language Recognition

Qi Wang<sup>1</sup>, Xilin Chen<sup>2</sup>, Chunli Wang<sup>2</sup>, and Wen Gao<sup>1,2</sup>

<sup>1</sup>*School of Computer Science and Technology, Harbin Institute of Technology, 150001, China*

<sup>2</sup>*Institute of Computing Technology, CAS, 100080, China*

{wangqi, xlchen, clwang, wgao}@jdl.ac.cn

## Abstract

*Viewpoint variance is one of the inevitable problems in vision based sign language recognition. However, most researchers avoid this problem by assuming a special view, especially the front view. In the paper, we propose a verification method for viewpoint invariant sign language recognition. In general, there are two major variances between two video sequences of the same sign: performance variance and viewpoint variance. For small performance variance, DTW can help us eliminate it. When there is only viewpoint variance between two sequences, we can consider the two sequences as obtained synchronously by a stereo vision system. Thus, for the current input, we can judge whether the known template is the matched one by verifying whether the two sequences can be considered as obtained by a stereo vision system. Our experiments demonstrate the efficiency of the proposed method. Furthermore, such verification method can be easily extended to other recognition tasks.*

## 1. Introduction

Sign language recognition is the typical case in computer science. First of all, sign language recognition has its significant realistic meaning. Because sign language is the primary modality of communication among deaf and mute society in the world, a working sign language recognition system would make deaf-hearing interaction easier. Furthermore, sign language recognition has its potential use in other applications such as Human-Computer Interaction. Moreover, a successful sign language recognition system will provide valuable insight into some other similar pattern recognition problems such as gait recognition, lip reading and human action identification.

Two major methods are used in the context of sign

language recognition [4]. They are data-glove based method and vision based method. Relatively speaking, vision based method is more natural and more convenient. However, there is a hard problem in vision based sign language recognition. It is the variance of viewpoint. Since the features will vary with the viewpoint, it is difficult for common methods, such as HMM [1, 3, 7], to apply to viewpoint invariant sign language recognition. The direct reason is the difficulty of extracting view invariant features. Thus, many researchers avoid the problem of viewpoint variance, by assuming a special view, especially the front view.

This paper addresses the problem of recognizing sign language in different viewpoint. As we know, few attentions focus on recognizing sign language in different viewpoint.

In [9], Wu and Huang proposed an appearance-based learning approach for view-independent recognition of static hand postures. However, the complexity of appearance in temporal signs makes the appearance-based learning method not fit for temporal sign language recognition.

For recognizing the temporal sign language in different viewpoint, Wang et al. [8] recently proposed a template matching method. In their opinion, two video sequences of the same sign could be roughly considered as obtained synchronously by a stereo vision system after time-warping and thus the fundamental matrix associated with two video sequences would be kept during all the period. So they proposed to recognize sign language in different viewpoint by verifying the uniqueness of the fundamental matrices, estimated from point correspondences at each instant in time. As well known, at least 8 point correspondences are needed for the estimation of fundamental matrix. We can see that, wang's method will suffer from the problem of possible lack of enough point correspondences. On one hand, it is possible that there are less than 8 point correspondences available at some instant in time. On

the other hand, lack of enough point correspondences will result in inaccurate estimation of the fundamental matrix, which will directly influence the efficiency of wang's method.

In the paper, we propose another verification method. Instead of estimating each fundamental matrix associated with each instant in time, we calculate only one fundamental matrix over all point correspondences between the given input and the warped template sequences. Obviously, the calculated fundamental matrix satisfies most of point correspondences only when the two sequences can be considered as obtained synchronously by a stereo vision system, which is just the case of the same sign. So we can judge whether the known template is the matched one by verifying whether the calculated fundamental matrix satisfies most of point correspondences. Since only one fundamental matrix will be estimated and all point correspondences will be used for the estimation, our method is more efficient and more robust than wang's method. Our experimental results demonstrate it.

The remainder of the paper is organized as follows: Section 2 gives the basic scheme of the proposed verification method. Section 3 describes its detail. Experimental results are presented in Section 4. Finally, the conclusion is given in Section 5.

## 2. The basic scheme

Considering that two sequences of the same sign can be roughly considered as obtained synchronously by a stereo vision system, we propose a verification method for viewpoint invariant sign language recognition. The basic scheme is depicted in Figure 1. Seen from the scheme, the verification methods involves two main steps: the first step is to warp the known template to the given input, so that small performance variance can be eliminated in the case of the same sign; the second step is to verify the existence of a stereo vision system where the two sequences can be considered as obtained synchronously, which is used to deal with viewpoint variance and judge the match for two sequences: matched if "exist" and unmatched if "not exist".

## 3. The verification method

The section will describe the detail of the proposed verification method.

### 3.1. Representing a sign

Within the literature of vision-based recognition of

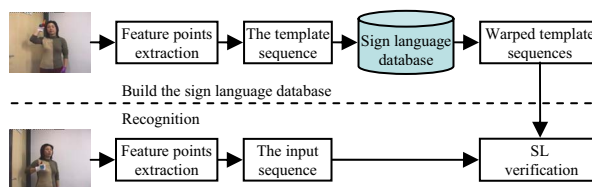


Figure 1. The scheme of the proposed verification method

sign language, a sign refers to a video clip. At a certain instant in time, the signer assumes a certain posture. If we use the concept of a configuration to denote the set of all visible feature points of a hand posture, written as  $C = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n\}$ , where  $\mathbf{p}_i = (x_i, y_i, 1)^T$  denotes a feature point and  $n$  is the number of feature points, we can represent a sign by  $S = \{C_1, C_2, \dots, C_t, \dots, C_m\}$ , where  $C_t$  is the configuration at time  $t$  and  $m$  is the length of the sequence.

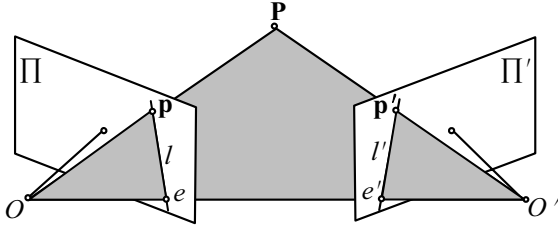
For feature points, we adopt color gloves and manually label them in our current works, as the focus of the paper is to recognize sign language using some points as feature, not to extract feature points in hands.

### 3.2. Verifying the Existence of a Stereo Vision System

As noted before, for a given input, we can judge whether the known template is the matched one by verifying whether the two sequences can be considered as obtained synchronously by a stereo vision system. In the section, we propose a two steps algorithm for the verification task.

The first step is to warp the template to the input, which is used to eliminate small performance variance in the case of the same sign. There, we adopt the same DTW technology as described in [6].

The second step is to verify whether the given input and the warped template can be considered as obtained synchronously by a stereo vision system. Two cases will happen. The first case is that there exists such a stereo vision system, which is the case of the same sign. In this case, a significant fundamental matrix can be calculated over all point correspondences between the given input and the warped template. The other case is that there is not such a stereo vision system, which is the case of the different signs. Obviously, this case also means that, there is no any fundamental matrix satisfies most of point correspondences between the two sequences. So we can see that, the task of verifying the existence of a stereo vision system is equivalent to the task of verifying whether the calculated fundamental matrix satisfies most of point correspondences between two sequences. There, we



**Figure 2.** The epipolar constraint:  $\mathbf{p}$  must lie in the epipolar line of  $l$  associated with  $\mathbf{p}'$  and  $\mathbf{p}'$  must lie in the epipolar line of  $l'$  associated with  $\mathbf{p}$  where  $\mathbf{p}$  and  $\mathbf{p}'$  are two images of a point  $\mathbf{P}$  observed by two cameras with optical centers  $O$  and  $O'$

propose a simple method to achieve the latter task, which involves the epipolar constraint [2] that, in any stereo vision system, an image point in one view must lie in the epipolar line associated with its corresponding point in the other view, as depicted in Figure 2.

From the knowledge of epipolar geometry, we can see that the epipolar constraint also mean that  $d(\mathbf{p}, \mathbf{F}\mathbf{p}') = 0$  where  $\mathbf{F}$  is the fundamental matrix associated with two views,  $\mathbf{F}\mathbf{p}'$  represent the epipolar line  $l$  associated with  $\mathbf{p}'$  and  $d(\mathbf{p}, l)$  denotes the spatial distance from the point  $\mathbf{p}$  to the line  $l$ . In the same manner, we can see that  $d(\mathbf{p}', \mathbf{F}^T\mathbf{p}) = 0$ .

Given  $n$  point correspondences and a fundamental matrix  $\mathbf{F}$ , we define

$$D(\mathbf{F}) = \frac{1}{n} \left( \sum_{i=1}^n \frac{1}{2} [d^2(\mathbf{p}_i, \mathbf{F}\mathbf{p}'_i) + d^2(\mathbf{p}'_i, \mathbf{F}^T\mathbf{p}_i)] \right) \quad (1)$$

From the epipolar constraint (as described above), we can see that,  $D(\mathbf{F})$  will be equal to 0 when  $\mathbf{F}$  satisfies all  $n$  point correspondences, while  $D(\mathbf{F})$  will be greater than 0 when  $\mathbf{F}$  doesn't satisfy all  $n$  point correspondences. We can also see that,  $D(\mathbf{F})$  will be close to 0 when  $\mathbf{F}$  satisfies most of  $n$  point correspondences, while  $D(\mathbf{F})$  will be well over 0 when  $\mathbf{F}$  satisfies only a small part of  $n$  point correspondences.

So we can use  $D(\mathbf{F})$  to judge whether the warped template is the matched one with the given input, where  $\mathbf{F}$  is the fundamental matrix calculated over all point correspondences between the two sequences. If  $D(\mathbf{F})$  is very closer to 0, we can see that  $\mathbf{F}$  satisfy most of point correspondences, which also mean that there exists a stereo vision system where the current input and the warped template can be considered as obtained synchronously. So, we can judge that the template is the matched one. In the similar manner, if  $D(\mathbf{F})$  is well over 0, we can judge that two sequences are not matched.

**Table 1.** The distance matrix for sign language recognition. The selected 6 CSL words are blues (1), hope (2), memory (3), difficulty (4), task (5), horse (6). The notation 1-T refers to 'the Template of sign word 1' and the notation 1-I refers to 'the input of sign word 1', etc. The distance value refers to the calculated  $D(\mathbf{F})$  according to Eq. (1). Note that lower values correspond to the two sequences of the same sign

	1-T	2-T	3-T	4-T	5-T	6-T
1-I	2.980	324.010	155.434	162.404	1898.621	1848.126
2-I	956.110	6.316	419.279	278.444	1036.500	859.799
3-I	1117.419	271.531	7.340	616.469	306.868	131.178
4-I	215.047	126.613	155.713	9.812	131.994	5683.55
5-I	165.417	481.693	118.996	119.626	9.394	78.284
6-I	1012.566	1742.900	789.013	348.179	314.479	7.578

Thus, we exploit the Nearest Neighbor rule to formalize the recognition task as follows:

$$T(S) = \underset{S \in \text{the template set}}{\operatorname{argmin}} D(\mathbf{F}^{S,S'}) \quad (2)$$

$$= \underset{S \in \text{the template set}}{\operatorname{argmin}} \frac{1}{n} \left( \sum_{i=1}^n \frac{1}{2} [d^2(\mathbf{p}_i, \mathbf{F}^{S,S'}\mathbf{p}'_i) + d^2(\mathbf{p}'_i, \mathbf{F}^{S,S'}\mathbf{p}_i)] \right),$$

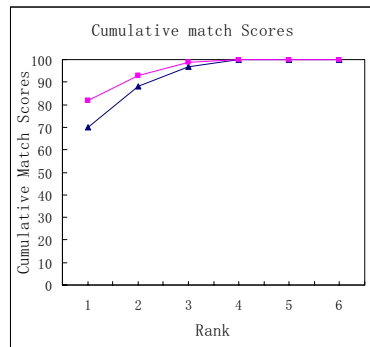
where  $n$  is the number of point correspondences between the given input and the warped template, and  $\mathbf{F}$  is the fundamental matrix calculated over all point correspondences between the two sequences.

## 4. Experimental results

Hindered by the difficulty of extracting view-invariant features, conventional methods, training a recognizer after extracting features, do not work well when the view is changed. The feasibility of the proposed verification method lies in that the method directly use all pairs of feature points between two sequences, which is viewpoint dependent, and thus bypass the difficulty of extracting view-invariant features.

To verify the proposed verification method for viewpoint invariant sign language recognition, we test it on a medium size vocabulary set (100 different signs). The data of template signs is collected from the frontal views and the data of input signs is collected from the views among  $0 \sim \pm 30^\circ$ . The input signs and the template signs are performed by the same signer at different speeds. For the boundary of each sign, we manually set the start point and the end point for now.

We illustrate the feasibility of the verification method in Table 1, where we select 6 sign words in CSL: blueness, hope, memory, difficulty, task and horse, and exhibit the value of  $D(\mathbf{F})$  between each input and each template. As expected, the value of  $D(\mathbf{F})$ , in



**Figure 3.** The CMS curves: the rectangle curve represents the recognition result of our verification method while the triangle curve represents the recognition result of wang's method [8]

the case of two sequences representing the same sign, is always lower than that in the case of two sequences representing the different signs.

For current collected data, the proposed verification method achieves a satisfying result. We show our performance evaluation in terms of cumulative match characteristics [5] in Figure 3. It can be seen that the recognition rate is 82% at rank 1, 93% at rank 2 and up to 99% at rank 3. Since only a fundamental matrix will be estimated and all point correspondences will be used for the estimation, our method is more efficient and more robust than wang's method [8], as depicted in Figure 3. The satisfying result demonstrates the efficiency of the proposed verification method, where the match of two sequences are judged by verifying whether they can be considered as obtained synchronously in a stereo vision system.

## 5. Conclusions

The paper has proposed a verification method for viewpoint-invariant sign language recognition with only one camera. There are two major steps in the verification method. The first step involves a DTW technology, which is used to eliminate small performance variance in the case of the same sign. The second step is verifying the existence of a stereo vision system where the given input and the warped template can be considered as obtained synchronously, which is used to deal with viewpoint variance and judge the match. Our experimental results demonstrated the efficiency of the proposed method. For a 100-word-vocabulary of Chinese Sign Language, the verification method achieved an accuracy of 82% at rank 1 and 93% at rank 2. Furthermore, the proposed method can be easily extended to other recognition task, such as gait recognition and lip-reading recognition.

Since great performance variance will bias the assumption that the two sequences of the same sign can be roughly considered as obtained synchronously by a stereo vision system, how to deal with the case will be investigated in the future. Some open problems, such as automatically seeking the start and end position within a continuous sign stream, will also be studied.

## Acknowledges

This research is supported by the National Science Council, R.O.China, under the Grant 60533030 and by Beijing Municipal Science Council, under the Grant 4061001.

## 6. References

- [1] B. Bauer and K.F. Kraiss, "Video-Based Sign Recognition Using Self-Organizing Subunits", International Conference on Pattern Recognition, 2002, pp. 434-437.
- [2] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2003.
- [3] R.M. McGuire, T. Starner, and et al., "Towards a one-way American sign language translator", International Conference on Automatic Face and Gesture Recognition, 2004, pp.620 – 625.
- [4] Sylvie C.W. Ong and Surendra Ranganath, "Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning", IEEE Trans. Pattern Analysis Machine Intelligence, 2005, 27(6): 873-891.
- [5] J. Phillips, H.Moon, S. Rizvi and P. Rause, "The FERET evaluation methodology for face recognition algorithms", IEEE Trans. Pattern Analysis and Machine Intelligence, 2000, 22: 1090-1104.
- [6] C. Rao, A. Gritai, M. Shah and T. Syeda-Mahmood, "View invariant alignment and matching of video sequences", International Conference on Computer Vision, 2003, pp. 939-945.
- [7] C. Vogler and D. Metaxas, "Asl recognition based on a coupling between hmms and 3d motion analysis", International Conference on Computer Vision, 1998, pp. 363-369.
- [8] Q. Wang, X. Chen, L. Zhang, C. Wang and W. Gao, "Viewpoint invariant sign language recognition". International Conference on Image Processing, 2005, pp. 281- 284.
- [9] Y. Wu and T. S. Huang, "View-independent recognition of hand postures", IEEE Conf. Computer Vision and Pattern Recognition, 2000, pp. 88-94.