Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 108 (2007) 87–97
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Viewpoint invariant sign language recognition

Qi Wang [a,*], Xilin Chen [b,*], Liang-Guo Zhang [b], Chunli Wang [b], Wen Gao [a,b]

[a] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China*
[b] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China*

Received 5 September 2005; accepted 21 November 2006
Available online 19 January 2007
Communicated by Mathias Kolsch

Abstract

Viewpoint invariance is a grand challenge for sign language recognition. In this paper, we propose a novel viewpoint invariant method for sign language recognition. The recognition task is converted to a verification task under the proposed method. This conversion is based on the geometric constraint that the fundamental matrix associated with two views SHOULD BE UNIQUE when the observation and template signs can be considered as obtained synchronously under a virtual stereo vision and vice versa. The Dempster–Shafer theory is applied to improve the robustness of the geometry model. Our experiment demonstrates the efficiency of the proposed method. Furthermore, the proposed method can be extended to other recognition tasks, such as gait recognition and lip-reading recognition.

Keywords: Sign language recognition; Viewpoint invariant recognition; Epipolar geometry; Fundamental matrix; Dempster–Shafer theory

1. Introduction

Sign language is the primary modality of communication among the hearing impaired. It is a kind of visual language via hand and arm movements accompanying facial expression and lip motion. The aim of sign language recognition is to provide an efficient and accurate mechanism to translate sign language into text or speech.

Sign language recognition will bridge communications between those with hearing impairment and those without. Meanwhile, this technology can also apply to the Human–Computer Interaction field. Inspired from sign language/gesture recognition, the same method can also be used to some other similar tasks, such as gait recognition, lip-reading recognition and human action classification.

The typical methods for sign language recognition, from the view of data input, include data-glove based method and vision-based method. The former can achieve a higher accuracy since the position, orientation, and angles are captured accurately. By contrast, a video camera is much cheaper and less constraining than data gloves. As a result, vision-based sign language recognition is receiving more attention.

There are many challenges posed by vision-based methods. For example, tracking the human hand is difficult because it is a complex non-rigid object. Modeling the dynamic aspect of sign language is complicated because it is very difficult to extract visual features that can effectively characterize handshape and motion. However, the grand challenge for sign language recognition is to seek a viewpoint invariant approach.

In this paper, we propose a novel method for viewpoint invariant sign language recognition. This method verifies the uniqueness for a virtual stereo vision system, which is formed by the observation and template. It utilizes the image coordinate of feature points and thus bypasses the difficulty of extracting view invariant features. Comparing with the other 3D model-based approaches, the advantage is that the proposed method can recognize the gesture directly, eliminating 3D reconstruction procedures.

* Corresponding authors.
E-mail addresses: wangqi@jdl.ac.cn (Q. Wang), xlchen@jdl.ac.cn (X. Chen).

1077-3142/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2006.11.009

Besides, it is a view-independent method, while the appearance-based method is sensitive to viewpoint change or a multiple appearance templates set is needed. The disadvantage is that this method depends on several pairs of feature points, which are hard to extract accurately from hands directly. Consequently colored gloves must be used.

The remainder of the paper is organized as follows: first, we give an overview of related work in Section 2. Then we describe the basic idea and details of the verification method for viewpoint invariant sign language recognition in Sections 3 and 4, respectively. We use the Dempster–Shafer method to improve the performance of verification in Section 5. We report our experimental results in Section 6 and conclude the paper in Section 7.

2. Related work

In the discussion of related work, we focus on previous efforts made in the area of understanding sign language by using the video camera. For coverage of gesture/sign language recognition, the surveys in [1,2] are an excellent starting point. Furthermore, some up-to-date works are reviewed in [3].

There are three main categories in sign language recognition: handshape classification, isolated sign language recognition and continuous sign language recognition.

Handshape classification, or fingerspelling recognition, is one of the main topics in sign language recognition since handshape can express not only some concepts, but also special transition states in temporal sign language. Triesch and Malsburg [4] apply the elastic graph matching technique to classify hand postures against complex backgrounds and achieve an accuracy of 86.2%. Bowden and Sarhadi [5] present a nonlinear model of shape and motion for tracking finger spelt American Sign Language (ASL). Lockton and Fitzgibbon [6] describe a real-time system for posture recognition, where real-time performance is provided by the combination of exemplar-based classification and a "deterministic boosting" algorithm which can allow for fast online retraining. Gupta and Ma [7] develop a complete vision-based system and reach 100% accuracy for 10 postures in ASL.

Isolated words are widely considered as the basic unit in sign language, and many researchers focus on isolated sign language recognition. Cui and Weng [8] use a recursive partition tree and apply Principal Component Analysis (PCA) and Multiple Discriminant Analysis (MDA) operations at each node to automatically derive the best subset features from images during learning. They achieve a recognition rate of 93.2%. Their method is able to achieve nonlinear classification boundaries in the feature space of 28 ASL signs.

Vogler and Metaxas [9] use three cameras placed in an orthogonal configuration to extract 3D parameters of a signer's body parts and utilize Hidden Markov Models (HMMs) for recognition. The accuracy rate is 89.9% for a vocabulary of 53 American Sign Language words.

Bauer and Kraiss [10,11] utilize subunits to enable scaling to large vocabulary. It achieves 92.5% accuracy over a lexicon of 100 words by using 150 HMM subunits. Encouragingly, they achieve a recognition rate of up to 81.0% for 50 new signs without retraining the subunit HMMs.

Bowden et al. [12] propose a linguistic feature vector for the visual interpretation of sign language. With this approach, they achieve a classification rate of up to 97.6% for a 43-word lexicon using only single instance for training.

Derpanis et al. [13] also investigate the representation and recognition of hand gesture within a linguistics-based framework. They demonstrate that the kinematic features recovered from the apparent motion can provide distinctive signatures for 14 primitive movements of ASL.

Some researchers also pay attention to continuous sign language recognition. A lot of works on continuous sign language recognition apply HMMs for recognition. The use of HMM offers the advantage of being able to segment a data stream into its continuous signs implicitly and thus bypasses the hard problem of segmentation entirely.

Starner and his colleagues [14,15] are the trailblazers of applying HMMs to sign language recognition. They use a view-based approach with a single camera to extract two dimensional features as the input of the HMMs and exploit a strongly constrained context model.

Bauer and Hienz [16] adopt continuous density HMMs and employ beam search to reduce computational complexity for the recognition task. Furthermore, they [17] apply stochastic language models to improve recognition performance.

Vogler and Metaxas [18] propose to use phonemes instead of whole signs as the basic units. In consideration of the smaller number of phonemes, it is possible to achieve larger vocabulary recognition.

However, few researchers consider viewpoint invariant sign language recognition. The reason may be the difficulty of extracting view invariant features. Many features used in conventional methods are sensitive to viewpoints. For example, the center-of-gravity of the hand blob [16,19,20], which is a widely used feature, highly depends on the viewpoint. Other features such as motion trajectories of hand pixels [21] also depend on the viewpoint where the sign is observed. Even if some in-plane rotation invariant features are used for gesture recognition [22], they cannot remain invariant for out-of-plane rotation.

Workable methods fall into two catalogues. The first is to use 3D methods. If a hand model can be reconstructed from a single image, we can achieve viewpoint invariant sign language recognition. However, the 3D reconstruction is time-consuming and sensitive to the configuration of the camera.

The other approach is to use appearance-based methods for recognition. Wu and Huang [23] propose an appearance-based learning approach for view invariant posture recognition. A small labeled data set and a large unlabeled data set are described through the same mixture density distribution, and a modified Discriminant-EM algorithm is used to estimate the parameters of the distribution. The recognition rate is 92.4% for 14 handshapes. This is an interesting step towards view invariant recognition. However, the appearance-based learning method is not suitable for temporal sign recognition, since the appearance of temporal signs will be protean under view variance besides temporal variance. Athitsos et al. [24] sample a hand model in discrete 3D orientations and measure the similarity between the observed appearance and discrete models. However, this method needs a huge storage space for models and consumes too much time matching observations to models.

In [25], Rao et al. propose to use the rank constraint of corresponding points in two views to establish temporal correspondence between the frames of two videos taken from different viewpoints. This idea inspired us to warp two sign sequences from different viewpoints and apply the method of verifying the uniqueness of fundamental matrices for viewpoint invariant sign language recognition.

3. Basic idea

The temporal–spatial sign language recognition process involves measuring the similarity between a current observation and one selected from a set of known observations. The paper presents a novel view for measuring the similarity between two sign sequences, which is independent to viewpoint change. Its basic idea is to consider the temporal–spatial sign language recognition task as a stereo vision task, where one camera gets the view of a known sign sequence while the other gets the view of the current observed one.

From the previous work on stereo vision [26,27], we know that the fundamental matrix of a stereo vision system should satisfy the constraint:

$$\mathbf{p}^{\mathrm{T}}\mathbf{F}\mathbf{p}' = 0, \tag{1}$$

where $\mathbf{p}$ and $\mathbf{p}'$ are the vectors of the same spatial point from the two cameras and $\mathbf{F}$ is the fundamental matrix. $\mathbf{F}$ can be regarded as the parameter of the stereo vision system. If the known sign sequence can be matched with the current observed one, then the fundamental matrix should be kept during all the periods of the sequences. The basic idea is illuminated in Fig. 1.

For two time-warped sequences which are performed for the same sign, they can be roughly considered as two sequences obtained synchronously from two views in a stereo vision system. Thus, the signers in the two sequences are assumed to hold the same posture for each moment and the fundamental matrices, which are derived from corresponding feature points at each moment, should be unique during the sign performed.

For two time-warped sequences which are performed for different signs, either the two synchronized postures cannot form a reasonable image pair of a stereo vision system, or the fundamental matrices are not unique.

Therefore, we can achieve the temporal–spatial sign language recognition task by verifying whether all estimated fundamental matrices are unique. We propose a two-step algorithm. First, we use the similar video synchronization technology as described in [25] to warp the known template to the current observation, where we assume that the template and observation represent the same sign. Next, we verify the uniqueness of fundamental matrices during the performance to determine whether they are the same sign or not.

Fig. 2. The colored gloves. (a) Palm of the dominant glove. (b) Back of the dominant glove. (c) Palm of the non-dominant glove. (d) Back of the non-dominant glove. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

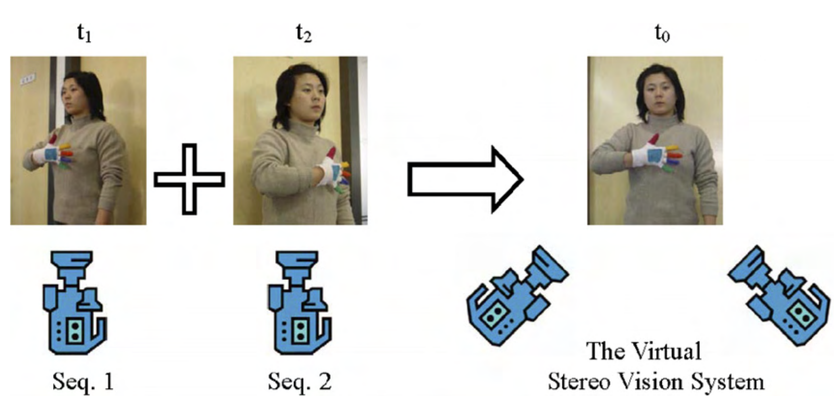

Fig. 1. The illumination of the basic idea: two time-warped sequences of the same sign can be roughly considered as the input of a stereo vision system, and the fundamental matrix associated with two views SHOULD BE UNIQUE.

4. Viewpoint invariant sign language recognition

This section will address three key issues pertaining to the proposed method for viewpoint invariant sign language recognition:

1. Sign representation
2. Alignment between a template and the observation sequence
3. Verification of the uniqueness of fundamental matrices.

4.1. Sign representation

In our work, feature points serve as the elements for sign representation. To track feature points easily and to describe the hand features more elaborately for discriminating the similar signs, a pair of colored cotton gloves is adopted. A well-designed color pattern can significantly simplify the tracking procedure. While performing sign language, the right hand typically plays a dominant role, and the left one plays the accessory or non-dominant role. We will refer to these as D-hand and ND-hand, respectively. Considering that the most discriminative feature information is conveyed by the D-hand, we color the D-hand glove with seven different colors to indicate the areas of the five fingers, palm and back. We paint the ND-hand glove with a single distinguishing color. The colored gloves are shown in Fig. 2. Such designation helps to not only extract feature points, but also find the matched features between consequent frames, such as corners of color regions and edges between different color regions.

We represent a sign by using the following method. A sign is a video clip. At a certain frame, the signer is in a certain posture. We represent the current posture by using the visible feature points and define as $C = \{\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n\}$, where $\mathbf{p}_i = (x_i, y_i, 1)^T$ denotes a feature point and n is the number of feature points. Then, we represent a sign as $S = \{C_1, C_2, \cdots, C_t, \cdots, C_m\}$, where C_t is the representation of hand posture at time t, and m is the length of the sequence.

4.2. Sign alignment

Accurately aligning two sequences of the same sign is vital to the success of the proposed method for viewpoint invariant sign language recognition, because it is only from the correct match that the fundamental matrix can be estimated properly. Rao et al. [25] propose a novel view-invariant DTW method to establish the temporal correspondence between frames of two videos representing the same action. In our work, we apply the similar method to align two sign sequences.

From the basic constraint of epipolar geometry, given two matched feature point sets:

$$C = \{\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n\}$$

$$C' = \{\mathbf{p}'_1, \mathbf{p}'_2, \cdots, \mathbf{p}'_n\}$$

a fundamental matrix $\mathbf{F}$ can be uniquely associated with (C, C') if the two point sets represent the same posture of a signer from two viewpoints. The resulting equation is as follows:

$$\mathbf{Af} = \begin{pmatrix} x_1x'_1 & \cdots & x_nx'_n \\ x_1y'_1 & \cdots & x_ny'_n \\ x_1 & \cdots & x_n \\ y_1x'_1 & \cdots & y_nx'_n \\ y_1y'_1 & \cdots & y_ny'_n \\ y_1 & \cdots & y_n \\ x'_1 & \cdots & x'_n \\ y'_1 & \cdots & y'_n \\ 1 & \cdots & 1 \end{pmatrix}^{\mathbf{T}} \begin{pmatrix} F_{11} \\ F_{12} \\ F_{13} \\ F_{21} \\ F_{22} \\ F_{23} \\ F_{31} \\ F_{32} \\ F_{33} \end{pmatrix} = 0, \tag{2}$$

where $\mathbf{f}$ is a 9-vector (in row-major order) representation of $\mathbf{F}$.

Since Eq. (2) is a homogenous equation, $\mathbf{A}$ has a rank of at most eight when the two point sets represent the same posture. On the other hand, the rank of $\mathbf{A}$ must be nine when the two point sets represent different postures. So, the ninth singular value of $\mathbf{A}^2$ can be used to measure the similarity of two hand postures [28]. Based on this similarity measurement, we perform DTW to warp each template to the current observation.

Intuitively, the accumulative distance after DTW can be used to measure the similarity of two sign sequences. However, our experiment suggests that the distance metric is not an efficient score to determine the degree of match between two sign sequences, as shown in Section 6.2. A new approach needs to be proposed for recognizing the sign from different viewpoints.

4.3. Verification of the uniqueness of fundamental matrices

As described before, we can recognize the sign from different viewpoints by verifying the uniqueness of fundamental matrices. The main steps are as follows: given the unknown observation S and the current Template S', we first warp S' to S as mentioned in Section 4.2. We note the unknown observation as $S = \{C_1, C_2, \cdots, C_m\}$ and the warped template as $S' = \{C'_1, C'_2, \cdots, C'_m\}$. Then, we obtain a sequence of the fundamental matrix, noted as $\mathbf{F}^{S,S'} = \{\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_m\}$, by analyzing point correspondences of each point-set pair (C_i, C'_i), where $i = 1, 2, \cdots, m$. Thereafter, we judge whether the unknown observation is the same sign as the known template by verifying the uniqueness of these fundamental matrices: "yes" if unique and "no" if not unique.

This section deals with the issue of verifying the uniqueness of fundamental matrices. A simple and efficient method is proposed. It utilizes the epipolar constraint that "an image point of $\mathbf{p}$ in one view must lie in the epipolar line of l associated with its corresponding point of $\mathbf{p}'$ in the other view" [26], as depicted in Fig. 3. Let us denote

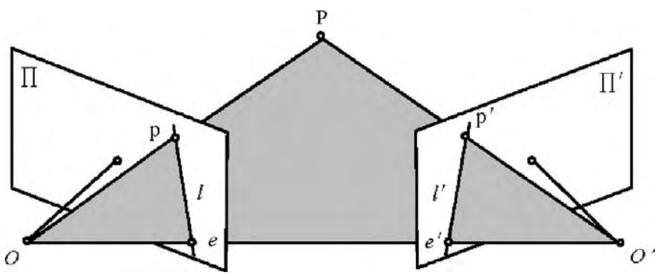

Fig. 3. The epipolar constraint: **p** must lie in the epipolar line of l associated with $\mathbf{p}'$ and $\mathbf{p}'$ must lie in the epipolar line of l' associated with **p** where **p** and $\mathbf{p}'$ are two images of a point **P** observed by two cameras with optical centers O and O'.

F as the fundamental matrix between the two views and $d(\mathbf{p}, l)$ as the spatial distance from the point **p** to the line l. Note that $\mathbf{Fp}'$ represents the epipolar line l associated with $\mathbf{p}'$, the epipolar constraint also means that $d(\mathbf{p}, \mathbf{Fp}')$ is equal to 0. In the same manner, $d(\mathbf{p}', \mathbf{F}^{\mathbf{T}}\mathbf{p})$ is equal to 0.

Given n point correspondences and a fundamental matrix **F**, we define

$$D(\mathbf{F}) = \frac{1}{n}\left(\sum_{i=1}^{n}\frac{1}{2}\left[d^2(\mathbf{p}_i, \mathbf{F}\mathbf{p}'_i) + d^2(\mathbf{p}'_i, \mathbf{F}^{\mathbf{T}}\mathbf{p}_i)\right]\right). \quad (3)$$

Lemma 1. $D(\mathbf{F}) = 0$ *if and only if all point correspondences are correct in the stereo vision systems represented by* **F**.

Remark. Using "the stereo vision systems" instead of "the stereo vision system", we refer to the universal set of all stereo vision systems that can be represented by the fundamental matrix **F**. It does not influence the correctness of Lemma 1, because one point correspondence will be correct in any stereo vision system of the universal set if they are correct in some stereo vision system of the universal set.

Proof. "if". This direction can be easily proved using the constraint that both $d^2(\mathbf{p}_i, \mathbf{F}\mathbf{p}'_i)$ and $d^2(\mathbf{p}'_i, \mathbf{F}^{\mathbf{T}}\mathbf{p}_i)$ will be equal to 0 if the point correspondence $(\mathbf{p}_i, \mathbf{p}'_i)$ is correct in the stereo vision systems represented by **F**.

"only if". This direction can be proved by reduction to absurdity. Suppose for contradiction that there exists one point correspondence $(\mathbf{p}_i, \mathbf{p}'_i)$ that is not correct in the stereo vision systems represented by **F** when $D(\mathbf{F}) = 0$. Following the epipolar constraint in Fig. 3, both $d^2(\mathbf{p}_i, \mathbf{F}\mathbf{p}'_i)$ and $d^2(\mathbf{p}'_i, \mathbf{F}^{\mathbf{T}}\mathbf{p}_i)$ will be greater than 0. It will directly give rise to $D(\mathbf{F}) > 0$, contradicting $D(\mathbf{F}) = 0$.

This completes the proof. □

Given $n > 8$ point correspondences and m fundamental matrices, we define

$$\begin{aligned} D &= \frac{1}{m}\sum_{t=1}^{m} D(\mathbf{F}_t) \\ &= \frac{1}{m}\sum_{t=1}^{m}\frac{1}{n}\left(\sum_{i=1}^{n}\frac{1}{2}\left[d^2(\mathbf{p}_i, \mathbf{F}_t\mathbf{p}'_i) + d^2(\mathbf{p}'_i, \mathbf{F}_t^{\mathbf{T}}\mathbf{p}_i)\right]\right). \end{aligned} \quad (4)$$

Lemma 2. $D = 0$ *if and only if all fundamental matrices are unique (up to scale) and all point correspondences are correct in the stereo vision systems represented by these fundamental matrices.*

Proof. Since the proof of the "if" part is obvious, we give only the proof of the "only if" part. By Lemma 1, we can say that for $\forall\ i, t$, $(\mathbf{p}_i, \mathbf{p}'_i)$ are corresponding in the stereo vision system represented by $\mathbf{F}_t$, otherwise, D must be greater than 0. Since the fundamental matrix can be uniquely determined by 8 point correspondences [26], we can conclude that all fundamental matrices are unique. This completes the proof. □

As mentioned before, we can roughly consider two sequences of the same sign as obtained synchronously by a stereo vision system. This fact implies that all estimated fundamental matrices will be unique in the case of the same sign. Based on Lemma 2, D will be equal to 0 in this case, where $\mathbf{F}_t$ is the estimated fundamental matrix at the tth instant in time. In a similar manner, D must be greater than 0 when two sequences represent different signs. So we can use the value of D to verify whether all estimated fundamental matrices are unique and thus judge whether two sequences represent the same sign.

However, many factors can cause D to be greater than 0 even in the case of the same sign, such as noise, the inaccurate location of feature points and misalignment between the observation and template. Even so, the value of D in the case of two sequences representing the same sign is far less than the value of D in the case of two sequences representing different signs. This fact brings us to the utilization of the Nearest Neighbor rule. We formalize our recognition task as follows:

$$\begin{aligned} T(S) &= \mathop{\arg\min}_{S' \in \text{the template set}} D^{S,S'} \\ &= \mathop{\arg\min}_{S' \in \text{the template set}} \frac{1}{m}\sum_{t=1}^{m}\frac{1}{n}\left(\sum_{i=1}^{n}\frac{1}{2}\left[d^2(\mathbf{p}_i, \overset{S,S'}{\mathbf{F}_t}\mathbf{p}'_i)\right.\right. \\ &\quad \left.\left. + d^2(\mathbf{p}'_i, \overset{S,S'}{\mathbf{F}_t^{\mathbf{T}}}\mathbf{p}_i)\right]\right), \end{aligned} \quad (5)$$

where m is the length of the unknown observation, n is the number of all point correspondences between the unknown one and the warped template, and $\mathbf{F}_t$ is the fundamental matrix estimated from point correspondences at the tth warped instant in time.

5. Towards a more robust measurement

There must be some instants in time with incorrect matches, because DTW cannot guarantee perfect alignment of two sequences. These instants may bring some negative influences to the verification framework in Eq. (5), as exhibited as follows:

We rewrite Eq. (4) by grouping all the matched points according to each instant in time as

$$D = \frac{1}{n} \times \frac{1}{m} \times \sum_{t=1}^{m} \left(\sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{1}{2} \left[d^2(\mathbf{p}_{ji}, \mathbf{F}_t \mathbf{p}'_{ji}) + d^2(\mathbf{p}'_{ji}, \mathbf{F}_t^{\mathrm{T}} \mathbf{p}_{ji}) \right] \right), \tag{6}$$

where n_j is the number of all point correspondences at the jth instant in time. Let

$$V(\mathbf{F}_t, j) = \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{1}{2} \left[d^2(\mathbf{p}_{ji}, \mathbf{F}_t \mathbf{p}'_{ji}) + d^2(\mathbf{p}'_{ji}, \mathbf{F}_t^{\mathrm{T}} \mathbf{p}_{ji}) \right], \tag{7}$$

Eq. (6) can be rewritten as

$$D = \frac{1}{n} \times \frac{1}{m} \times \sum_{t=1}^{m} \sum_{j=1}^{m} n_j V(\mathbf{F}_t, j) = \frac{1}{n} \sum_{j=1}^{m} \left(\frac{1}{m} \sum_{t=1}^{m} n_j V(\mathbf{F}_t, j) \right).$$

Let $D_j = \frac{1}{m} \sum_{t=1}^{m} n_j V(\mathbf{F}_t, j)$, we have

$$D = \frac{1}{n} \sum_{j=1}^{m} D_j. \tag{8}$$

This definition indicates that D_j represents a partial verification result for the uniqueness of all fundamental matrices. Let us to see what happens if the match at the jth instant in time is not correct. $V(\mathbf{F}_t, j)$ (Eq. 7) will be well over 0. Consequently, D_j will be well over 0, which directly results in D (Eq. 8) being well over 0. This suggests that those instants with incorrect matches may bring some negative influences, because the verification framework in Eq. (5) is based on the fact that D will be close to 0 when two sequences represent the same sign.

This section presents an improved version of the verification framework to weaken the negative influence of badly matched instants in time and strengthen the contribution of accurately matched instants in time. In the improved version, we explain the verification task as a decision problem in the framework of the Dempster–Shafer theory.

5.1. Dempster–Shafer theory

A decision problem is composed of a set of mutually exclusive hypotheses called *frame of discernment* and noted as Ω [29]. Establishing the correct hypothesis depends on information that the evidence provides. Information that contributes to the knowledge of the problem is captured by a mass function:

$$m : 2^{\Omega} \to [0,1],\ m(\phi) = 0,\ \sum_{B \subseteq \Omega} m(B) = 1 \tag{9}$$

or a belief function

$$\begin{cases} Bel : 2^{\Omega} \to [0,1],\ Bel(\phi) = 0,\ Bel(\Omega) = 1, \\ \text{and for any collection } A_1, A_2, \cdots, A_n \\ Bel(A_1 \cup \cdots \cup A_n) \geqslant \sum\limits_{\substack{I \subseteq \{1 \ldots n\} \\ I \neq \phi}} (-1)^{|I|+1} Bel\left(\bigcap\limits_{i \in I} A_i \right). \end{cases} \tag{10}$$

In fact, the two representations are equivalent. They are related by the following two equations

$$\begin{cases} Bel(A) = \sum\limits_{B \subseteq A} m(B), \text{ for all } A \subseteq \Omega \\ m(A) = \sum\limits_{B \subseteq A} (-1)^{|A-B|} Bel(B), \text{ for all } A \subseteq \Omega. \end{cases} \tag{11}$$

For two distinct evidences, the Dempster rule is used to combine them according to

$$\begin{cases} m_{12}(A) = (m_1 \oplus m_2)(A) = K^{-1} \sum\limits_{A_1 \cap A_2 = A} m_1(A_1) m_2(A_2), \\ K = 1 - \sum\limits_{A_1 \cap A_2 = \phi} m_1(A_1) m_2(A_2), \end{cases} \tag{12}$$

where K indicates the degree of conflict between the sources.

5.2. Applying Dempster–Shafer theory to the verification task

A verification task can be considered to be a decision problem. This section involves applying Dempster–Shafer theory to formalize the verification task of the uniqueness of fundamental matrices as a decision problem.

One of the keys to a decision problem is to search for evidences, which provide information to the knowledge on this problem. By Lemmas 1 and 2, point correspondences can be used to verify whether a set of fundamental matrices are unique. We define that

$$I_j = \frac{1}{m} \sum_{t=1}^{m} V(\mathbf{F}_t, j) = \frac{1}{m} \sum_{t=1}^{m} \left(\frac{1}{n_j} \sum_{i=1}^{n_j} \frac{1}{2} \left[d^2(\mathbf{p}_{ji}, \mathbf{F}_t \mathbf{p}'_{ji}) + d^2(\mathbf{p}'_{ji}, \mathbf{F}_t^{\mathrm{T}} \mathbf{p}_{ji}) \right] \right) \tag{13}$$

where j represents the jth instant in time, m is the length of the unknown observation, $\mathbf{F}_t$ is the fundamental matrix estimated from point correspondences at the tth instant in time, $V(\mathbf{F}_t, j)$ is defined in Eq. (7), n_j is the number of point pairs at the jth instant in time.

The equation of Eq. (7) shows that $V(\mathbf{F}_t, j)$ will be equal to 0 only when $\mathbf{F}_t$ is the exact fundamental matrix associated with the jth instant in time. It suggests that I_j will be equal to 0 only when all fundamental matrices are the exact fundamental matrix associated with the jth instant in time. That is, I_j conveys some information whether these fundamental matrices are unique or not. In fact, I_j provides the verification result for the uniqueness of these fundamental matrices when using only point correspondences at the jth instant in time.

As mentioned before, there must be some instants in time with incorrect matches even for two sequences of the same sign. What happens to I if there are some incorrect matches? Its definition (Eq. 13) shows that, in the case of two sequences representing the same sign, I_{k_1} will be less than I_{k_2} if the match at the k_1th instant in time is correct and the match at the k_2th instant in time is incorrect. This suggests that I implicitly reflect the match degree at the associated instant in time in the case of the same sign.

Writing all I_j in increasing order as $I^{S,S'} = \{I_{(1)}, I_{(2)}, \cdots, I_{(m)}\}$, we define that

$$\begin{cases} E_1 = \frac{1}{n_1}\sum\limits_{k=1}^{n_1} I_{(k)},\ n_1 = \lceil mq_1 \rceil,\ 0 < q_1 < 1, \\ E_2 = \frac{1}{n_2}\sum\limits_{k=m-n_2+1}^{m} I_{(k)},\ n_2 = \lceil mq_2 \rceil,\ 0 < q_2 < 1, \end{cases} \tag{14}$$

where q_1 and q_2 are two constants. This definition indicates that E_1 identifies the top $\lceil mq_1 \rceil$ best matched frames when S is the same sign as S' and E_2 reflects the unmatched degree when S is distinct from S'. Further experiments show that there is such a relationship as depicted in Fig. 4, where the X axis represents the degree of match between two sign clips and the Y axis represents the value of E_1 or E_2. As seen from Fig. 4, E_1 will be little if S is the same sign as S' and that E_2 will be great if S is distinct from S'. So, we are reasonable to believe that S is the same sign as S' when E_1 is very little and that S is distinct from S' when E_2 is very great.

Thus, we consider E_1 and E_2 as two evidences. By applying the Dempster–Shafer theory, we formalize the verification task of the uniqueness of fundamental matrices as a decision problem:

1. The *frame of discernment* Ω is

$$\Omega = \{\textit{uniqueness},\ \textit{non-uniqueness}\}, \tag{15}$$

note $A = \{\textit{uniqueness}\}$,
$B = \{\textit{non-uniqueness}\}$;

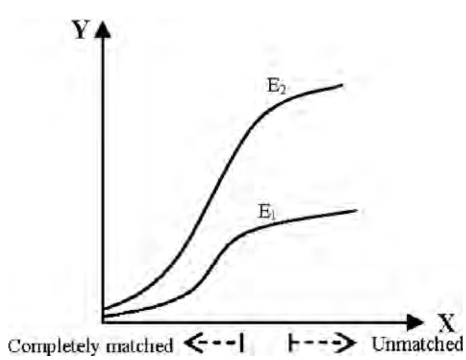

Fig. 4. The relationship curve between the value of E_1 or E_2 and the degree of match between two sign clips.

2. The belief function for the evidence of E_1 is

$$\begin{cases} Bel_1(A) = \max\left\{c_1(1 - \frac{E_1}{C_1}), 0\right\}, \\ Bel_1(B) = 0, \\ Bel_1(\Omega) = 1; \end{cases} \tag{16}$$

3. The belief function for the evidence of E_2 is

$$\begin{cases} Bel_2(A) = 0, \\ Bel_2(B) = \min\left\{c_2\frac{E_2}{C_2}, 1\right\}, \\ Bel_2(\Omega) = 1, \end{cases} \tag{17}$$

where c_1, c_2 are two constants which could be fine-tuned by experiment, here we set $c_1 = 1$ and $c_2 = 5$; C_1 is a constant with the value of 2048 and C_2 is a constant with the value of $320/2 \times 320/2$, where 320 is the width of the image. Two factors are considered to set these constants: the first is that we hope $Bel_1(A)$ is close to 1 when E_1 is little; the second is that we hope $Bel_2(B)$ is close to 1 when E_2 is greater than 70×70.

According to Eq. (11), we can obtain the mass function for the evidence of E_1 as:

$$\begin{cases} m_1(A) = \max\left\{c_1(1 - \frac{E_1}{C_1}), 0\right\}, \\ m_1(B) = 0, \\ m_1(\Omega) = 1 - \max\left\{c_1(1 - \frac{E_1}{C_1}), 0\right\} \end{cases} \tag{18}$$

and the mass function for the evidence of E_2 as:

$$\begin{cases} m_2(A) = 0, \\ m_2(B) = \min\left\{c_2\frac{E_2}{C_2}, 1\right\}, \\ m_2(\Omega) = 1 - \min\left\{c_2\frac{E_2}{C_2}, 1\right\}. \end{cases} \tag{19}$$

By applying the Dempster rule (Eq. (12)) and according to Eq. (11), we can combine the two evidences of E_1 and E_2 as:

$$\begin{cases} Bel_{12}(A) = K^{-1} \times \max\left\{c_1(1 - \frac{E_1}{C_1}), 0\right\} \\ \qquad \times \left(1 - \min\left\{c_2\frac{E_2}{C_2}, 1\right\}\right), \\ Bel_{12}(B) = K^{-1} \times \left(1 - \max\left\{c_1(1 - \frac{E_1}{C_1}), 0\right\}\right) \\ \qquad \times \min\left\{c_2\frac{E_2}{C_2}, 1\right\}, \\ Bel_{12}(\Omega) = 1, \\ \text{where } K = 1 - \max\left\{c_1(1 - \frac{E_1}{C_1}), 0\right\} \\ \qquad \times \min\left\{c_2\frac{E_2}{C_2}, 1\right\}. \end{cases} \tag{20}$$

We consider a template S' as the matched one when we have the highest degree of belief of the uniqueness of fundamental matrices for S' among all templates. So we formalize the recognition task by applying Dempster–Shafer theory as:

$$
\begin{aligned}
T(S) &= \mathop{\arg\max}_{S' \in \text{the template set}} \left(Bel_{12}^{S,S'}(A)\right) \\
&= \mathop{\arg\max}_{S' \in \text{the template set}} K^{-1} \times \max\left\{c_1\left(1-\frac{E_1}{C_1}\right), 0\right\} \\
&\quad \times \left(1-\min\left\{c_2\frac{E_2}{C_2}, 1\right\}\right).
\end{aligned} \tag{21}
$$

6. Experimental results

To evaluate the proposed verification method for viewpoint invariant sign language recognition, we test it on a medium size vocabulary set (100 different signs in Chinese Sign Language, as shown in Appendix A). All Signs are performed by a native signer and collected by a USB color camera in the laboratory environment with unconstrained background and fluorescent illumination. The image revolution is 320×240. Each sign has five samples together, where four samples are collected from the frontal view and one sample is collected from the non-frontal view. Most non-frontal-view samples are collected from the views of $\pm 30°$. We use one of four samples from the frontal view as the template sequence and the only sample from the non-frontal view as the test sequences for the experiments in Sections 6.2 and 6.3.

We first show that it is difficult for HMMs to apply to the task of viewpoint invariant sign language recognition. Then we experiment with the accumulative distance in DTW, as mentioned in Section 4.2, and show that the distances measurement is not an efficient score for determining the degree of match between two sequences. Finally, we demonstrate the efficiency of the proposed verification method at the end of this section.

6.1. HMMs

We design three experiments, Experiment 1, Experiment 2 and Experiment 3, to show that HMMs are not suitable for recognizing the signs from different views (the adopted HMM is from Zhang [30]). In Experiment 1, we use samples from the same view for the training data and the test data. In Experiment 2, we train the HMM with samples from one view and test it with samples from the other view. In Experiment 3, we train HMMs with samples from more than one view. The experimental results are shown in Table 1.

The experimental results suggest that: the HMM is an efficient method when the view of the test data is the same as that of the training data, as Experiment 1 shows. The HMM trained with sign sequences in one view cannot work well with sign sequences in another view, as the result of Experiment 2 indicates. The HMM trained with sign sequences in more than one view also cannot work well, as the result of Experiment 3 suggests. We ascribe the experimental results to the fact that the used feature [30] changes with the view.

The difficulty of extracting view invariant features hinders HMM from being applied to view invariant sign language recognition.

6.2. The accumulative distance in DTW

It can be assumed that the accumulative distance in DTW, as mentioned in Section 4.2, can be used to measure the similarity of two sequences, since the ninth singular value of $\mathbf{A}^2$ can be used to measure the similarity of two postures. However, our experiment shows that the distance measurement is not an efficient score for determining the degree of match between two sequences. The experimental result is shown in Fig. 5 in terms of cumulative match characteristics [31].

We think the reason for the low recognition rate may be that the ninth singular value can only give an answer to whether two hand postures are matched or not matched. It is possible that the ninth singular value is able to reflect the degree of match in the case of similar hand postures, i.e., these postures around the corresponding instant in time between two sequences of the same sign. So the ninth singular value can be used as a similarity measurement to warp the two sequences of the same sign. However, it is doubted that the ninth singular value is able to reflect the unmatched degree in the case of different hand postures, because the fact that the ninth singular value is not equal to 0 can only infer that the two hand postures do not match.

So we need to seek another efficient measurement.

6.3. The verification method

For current collected data, the proposed verification method achieves a satisfying result. We show our performance evaluation in Fig. 6, as the curve represented by

Table 1
HMMs: results of the recognition experiment

Experiment	Training data	Test data	Details	Accuracy (%)
Experiment 1	3 S. in F	1 S. in F	$M = 3$, $N = 3$	94.0
Experiment 2	3 S. in F	1 S. in NF	$M = 3$, $N = 3$	41.0
Experiment 3	2 S. in F and 1 S. in NF	1 S. in F	$M = 3$, $N = 3$	57.0

Note. The experiment exploits Zhang's technology [30] and uses a 100-word-vocabulary of Chinese Sign Language. The notation '3 S. in F' refers to '3 Samples in Frontal view' and The notation '1 S. in NF' refers to '1 Sample in Non-Frontal view'. Experiment 1 shows that HMM is an efficient method for the case of the same viewpoint, Experiment 2 shows that the HMM trained with the samples from one viewpoint cannot work well with the sequences from another viewpoint, and Experiment 3 shows that it is not appropriate to train the HMM with samples from more than one viewpoint.

M denotes the number of mixture-components.

N denotes the number of states.

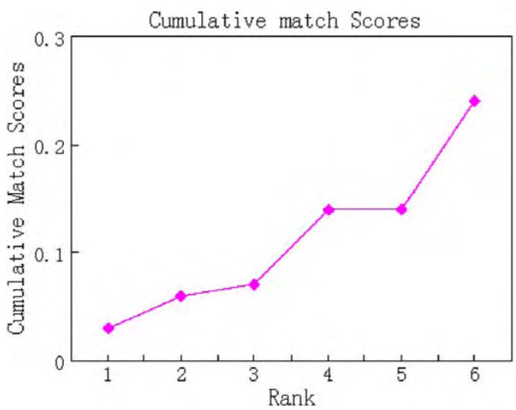

Fig. 5. The CMS curves for the accumulative distance in DTW.

rectangles shows. The recognition rate is 79.0% at rank 1 and reaches 92.0% at rank 2. The satisfying result demonstrates the efficiency of considering the temporal–spatial recognition task as a verification task within a stereo vision framework.

We further illustrate the proposed verification method in Fig. 7. The unknown observation is shown in the fourth row, which is the word ''glory'' obtained from the view of −20°. Two template signs include ''glory'' and ''sky'' obtained from the front view, which are shown in the second row and the sixth row, respectively. We first exhibit the verification process for the template sign ''glory''. The verification process begins with warping the template sign to the unknown observation, as described in Section 4.2. We show the warped results in the second row and the fourth row, where the columns represent the warped corresponding frames. From point correspondences, we estimate the associated fundamental matrices at each warped instant in time. Then, we calculate the verification information I for each warped instant in time according to Eq. (13). The calculated information for selected time instances is shown in the third row. Next, we compute the evidences of E_1 and E_2 according to Eq. (14), where q_1 and q_2 are set to 0.3. Afterwards, by applying the Dempster–Shafer theory as described in Section 5, we obtain the belief degree of the uniqueness of fundamental matrices for the template sign ''glory'', which is 0.93. In the same manner, we obtain the belief degree for the template sign ''sky''. The warped result and the calculated information for selected time instances are exhibited in 4–6 rows. The obtained belief degree is 0.28. Finally, we recognize the unknown observation as the template sign with the maximum belief degree. Here, it is the template sign ''glory''.

We also report the results with individual similarity measurements of D (Eq. 5), E_1 and E_2 (Eq. 14) in Fig. 6. The ''a'' figure serves to compare the recognition performance of using the similarity measurement of D with the recognition performance of applying Dempster–Shafer theory. It shows that the performance has been improved by 9% points when Dempster–Shafer theory is applied. The improved performance suggests that it is efficient to make a difference between the accurately matched instants in time and the badly matched instants in time. It also suggests that it is efficient to formalize the verification task as a decision problem in the framework of the Dempster–Shafer theory. The ''b'' figure gives the result when using individually E_1 or E_2 as the similarity measurement between two signs. This figure shows that E_1 and E_2 provide evidence for the degree that two sequences are matched or not matched, since comparative recognition rates relative to D have been obtained. At the same time, improved performance is achieved when Dempster–Shafer theory is applied (7% points over E_1 and 8% points over E_2). The improved performance does indicate the efficiency of applying the Dempster rule to combine both evidences.

To verify the robustness of the proposed method, we collect five new test sets from the view of 0°, 5°, 30°, 60° and 75°, respectively. The experimental results are shown in Fig. 8. It can be found that the accuracy increases when

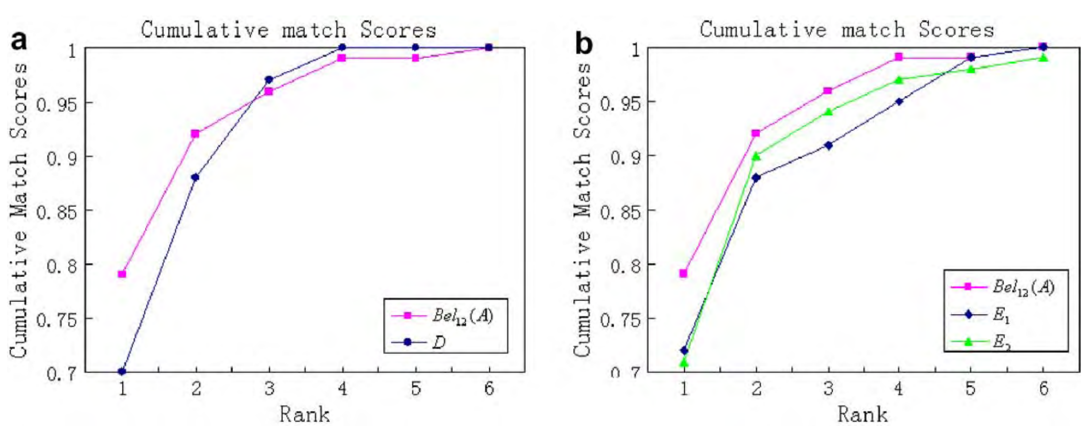

Fig. 6. The CMS Curves: the curve represented by rectangles illustrates the performance of applying Dempster–Shafer theory to the verification task, the curve represented by circles illustrates the performance of the verification framework using D as the measurement, the curve represented by diamonds and the curve represented by triangles illustrate the performance of using individually E_1 and E_2 as the measurement, respectively. The left figure (a) demonstrates the efficiency of making a difference between the accurately matched instants in time and the badly matched instants in time, and the efficiency of formalizing the verification task as a decision problem in the framework of Dempster–Shafer theory. The right figure (b) demonstrates the efficiency of applying the Dempster rule to combine both evidences.

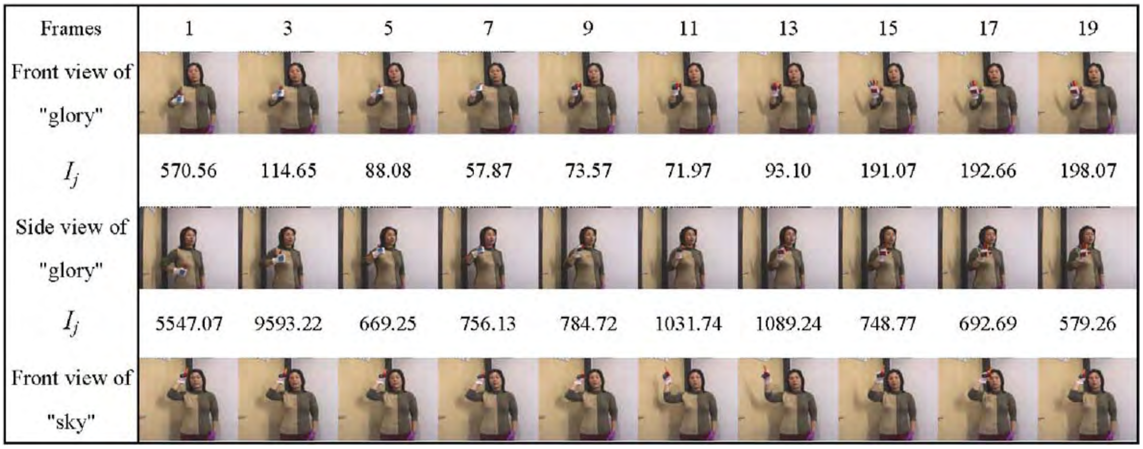

Fig. 7. The illumination of applying the proposed verification framework to recognize the signs from different viewpoint: the unknown observation is the word of "glory" as shown in the fourth row, template 1 is the same sign of "glory" from different viewpoint as shown in the second row and template 2 is the word of "sky" as shown in the sixth row. The columns represent the warped corresponding frames. Every 2nd output frame is shown. They are frames 1, 3, 5, 7, 9, 11, 13, 15, 17, 19. The third and fifth rows show the calculated verification information of I at each corresponding instant in time for template 1 and template 2, respectively.

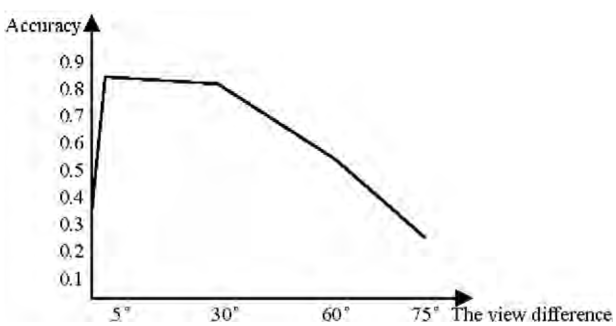

Fig. 8. The effect of viewpoint changes on performance of the proposed verification method.

the difference of view angles between observing and template modeling decreases, except when the difference reduces to 0°. A very low recognition rate is achieved when the view difference is close to 0°. The reason is that we cannot get the fundamental matrix for verification in this case. When the difference of the view direction is limited to the scope of 30°, the proposed method keeps a high recognition rate since there are enough point correspondences. When the difference of the view angle is greater than 30°, the accuracy of feature point matching decreases. Consequently, the estimation of fundamental matrices will be sensitive. This directly leads to the decrease of the recognition accuracy. Although the proposed method cannot achieve a high recognition rate in all cases, this experiment indicates that the proposed verification method can work well when the viewpoint changes in a limited range.

7. Conclusions

The paper presents a viewpoint invariant sign language recognition approach with only one camera. The proposed method converts the temporal–spatial recognition task as a verification task of a virtual stereo vision system. The Dempster–Shafer method is used to improve the recognition accuracy. For a 100-word-vocabulary of Chinese Sign Language, the proposed method achieves an accuracy of 92% at rank 2. Furthermore, the proposed method could also be extended to other similar tasks, such as gait recognition and lip-reading recognition.

Although the proposed verification method avoids the viewpoint-dependent problem, several issues can enhance this verification framework. Under the current framework, the verification will fail when the unknown observation and the matched template are captured from an identical viewpoint. In that case, we cannot get the fundamental matrix for verification. This can be solved by providing two templates from different viewpoints for each sign. Besides, incorporating multiple templates is necessary since the performance of the same sign varies from time to time even for the same signer. As a next step, we will apply a statistical method to measure the differences between the observation and templates. Typically, the Linear Discriminant Analysis will be utilized. Meanwhile, automatic segmentation of the sign from the video stream is also a key issue in our future work.

Acknowledgments

This research is partially supported by Natural Science Foundation of China under contract Nos. 60533030 and 60603023, and also by Natural Science Foundation of Beijing Municipal under contract No. 4061001. The authors thank Miss Yan Ma, who provides great help on data collection. The authors also thank Isaac Harris from Carnegie Mellon University for proofreading this manuscript.

Appendix A. The vocabulary for testing

absorb	admit	afternoon	aged	alter	always	Argentina	ash	Asia	assignment
be deceived	be well	beauty	before	believe	black	blue	blueness	breath out	bright
broadcast	Canada	category	chairman	change	chatter	Chengdu	chicken	China	color
corn	cry	daytime	demand	die hard	difficulty	dirty	effort	endeavour	everybody
expectation	face	fancy	fearfulness	forget	fragrant	furious	get rid of	glory	good bye
grief	have	heat	hope	horse	hot	however	ideal	insect	question
interpret	invent	Korea	lean	loneliness	maze	memory	Mexico	mind	Mongolia
morning	nature	news	Norway	offensive	officer	often	pig	power	regime
report	sister	sky	snake	stimulate	tasteless	Thailand	them	theory	thought
tomorrow	toothbrush	Ukraine	veins	Washington	we	who	you	youth	Zhang

References

[1] V.I. Pavlovic, R. Sharma, T.S. Huang, Visual interpretation of hand gestures for human–computer interaction: a Review, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 677–695.

[2] Y. Wu, T. Huang,Vision-based gesture recognition: a review, in: A. Braffort, R. Gherbi, S. Gibet, J. Richardson, D. Teil (Eds.), Gesture-Based Communication in Human–Computer Interaction, Lecture Notes in Artificial Intelligence, vol. 1739, Springer-Verlag, Berlin, 1999, pp. 103–115.

[3] S.C.W. Ong, S. Ranganath, Automatic sign language analysis: a survey and the future beyond lexical meaning, IEEE Trans. Pattern Anal. Mach. Intell. 27 (6) (2005) 873–891.

[4] J. Triesch, C. Malsburg, Robust classification of hand postures against complex background, in: Intl Conf. on Automatic Face and Gesture Recognition, 1996.

[5] R. Bowden, M. Sarhadi, A non-linear model of shape and motion for tracking finger spelt american sign language, Image Vis. Comput. 20 (9-10) (2002) 597–607.

[6] R. Lockton, A.W. Fitzgibbon, Real-time gesture recognition using deterministic boosting, in: British Machine Vision Conf., 2002.

[7] L. Gupta, S. Ma, Gesture-based interaction and communication: automated classification of hand gesture contours, IEEE Trans. Systems, Man, and Cybernetics, Part C: Application Rev 31 (1) (2001) 114–120.

[8] Y. Cui, J. Weng, Appearance-based hand sign recognition from intensity image sequences, Comput. Vis. Image Understand. 78 (2) (2000) 157–176.

[9] C. Vogler, D. Metaxas, ASL recognition based on a coupling between hmms and 3d motion analysis, in: Intl. Conf. on Computer Vision, 1998, pp. 363–369.

[10] B. Bauer, K.F. Kraiss, Towards an automatic sign language recognition system using subunits, in: Gesture Workshop, 2001, pp. 64–75.

[11] B. Bauer, K.F. Kraiss, Video-based sign recognition using self-organizing subunits, in: Intl. Conf. on Pattern Recognition, 2002, pp. 434–437.

[12] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, M. Brady, A linguistic feature vector for the visual interpretation of sign language, in: Euro. Conf. on Computer Vision, 2004, pp. 390–401.

[13] K.G. Derpanis, R.R. Wildes, J.K. Tsotsos, Hand gesture recognition within a linguistics-based framework, in: Euro. Conf. on Computer Vision, 2004, pp. 391–401.

[14] T. Starner, A. Pentland, Visual recognition of american sign language using hidden markov models, in: Intl. Conf. on Automatic Face and Gesture Recognition, 1995, pp. 189–194.

[15] T. Starner, J. Weaver, A. Pentland, Real-time american sign language recognition using desk and wearable computer based video, IEEE Trans. Pattern Anal. Mach. Intell. 20 (12) (1998) 1371–1375.

[16] B. Bauer, H. Hienz, Relevant features for video-based continuous sign language recognition, in: Intl. Conf. on Automatic face and Gesture recognition, 2000, pp. 440–445.

[17] B. Bauer, H. Hienz, K.F. Kraiss, Video-based sign language recognition using statistical methods, in: Intl. Conf. on Pattern Recognition, 2000, pp. 463–466.

[18] C. Vogler, D. Metaxas, A framework for recognizing the simultaneous aspects of american sign language, Comput. Vis. Image Understand. 81 (2001) 358–384.

[19] M. Assan, K. Grobel, Video-based sign language recognition using hidden markov models, in: Gesture Workshop, 1997, pp. 97–109.

[20] N. Tanibata, N. Shimada, Y. Shirai, Extraction of hand features for recognition of sign language words, in: Intl. Conf. on Vision Interface, 2002, pp. 391–398.

[21] M.H. Yang, N. Ahuja, M. Tabb, Extraction of 2D motion trajectories and its application to hand gesture recognition, IEEE Trans. Pattern Anal. Mach. Intell. 24 (8) (2002) 1061–1074.

[22] O. Al-Jarrah, A. Halawani, Recognition of gestures in arabic sign language using Neuro-Fuzzy systems, Artif. Intell. 133 (2001) 117–138.

[23] Y. Wu, T.S. Huang, View-independent recognition of hand postures, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2000, vol. 2, pp. 88–94.

[24] V. Athitsos, J. Alon, S. Sclaroff, G. Kollios, Boostmap: a method for efficient approximate similarity rankings, in: IEEE Conf. on Computer Vision and Pattern Recognition, 2004, pp. 268–275.

[25] C. Rao, A. Gritai, M. Shah, T. Syeda-Mahmood, View invariant alignment and matching of video sequences, in: Intl. Conf. on Computer Vision, 2003, pp. 939–945.

[26] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2003.

[27] Q.T. Luong, O.D. Faugeras, Self calibration of a moving camera from point correspondences and fundamental matrices, IJCV 22 (3) (1997) 261–289.

[28] A. Gritai, Y. Sheikh, Mubarak Shah, On the use of anthropometry in the invariant analysis of human actions, in: Intl. Conf. on Pattern Recognition, 2004, pp. 923–926.

[29] G. Shafer, A Mathematical Theory of Evidence, P.U. Press, Princeton, New Jersey, 1976.

[30] L. Zhang, Y. Chen, G. Fang, X. Chen, W. Gao, A vision-based sign language recognition system using tied-mixture density HMM, in: Intl. Conf. on Multimodal Interfaces, 2004, pp.198–204.

[31] J. Phillips, H. Moon, S. Rizvi, P. Rause, The FERET evaluation methodology for face recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 1090–1104.