SIGN LANGUAGE RECOGNITION FROM HOMOGRAPHY

Qi Wang[1], Xilin Chen[2], Chunli Wang[2], and Wen Gao[1,2]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China
[2]Institute of Computing Technology, CAS, 100080, China
{wangqi, xlchen, clwang, wgao}@jdl.ac.cn

ABSTRACT

It is difficult to recognize sign language in different viewpoint. The HMM method is hindered by the difficulty of extracting view invariant features. The general template matching methods have a strong constraint such as accurate alignment between the template sign and the test sign. In the paper, we introduce a novel approach for viewpoint invariant sign language recognition. The proposed approach requires no view invariant features, low training and no alignment. Its basic idea is to consider a sign as a series of tiny hand motions and utilize the HOMOGRAPHY of tiny hand motions. Using the word of "homography", we mean that there are the same tiny hand motions as well as their appearance order in different performances of the same sign. The experimental results demonstrate the efficiency of the proposed method.

1. INTRODUCTION

Vision based sign language recognition is the typical case in computer vision and has all along attracted researcher's attention [9].

As early as 1995, Starner and Pentland[12] did study continuous sign language recognition. Their later works [13, 7] had an intent to actualize a mobile one-way American Sign Language translator. Vogler and Metaxas experimented with 3-D camera system and obtained an 89.9% accuracy for a 53-word lexicon [15]. Besides, Vogler [16] challenge the problem of finding a modeling paradigm that would scale well with increasing vocabulary size. They proposed to use phonemes as the smallest identifiable subunits. Since the number of phonemes was limited, such method was scalable. The same problem was addressed by Bauer and Kraiss [2], who utilized self-organizing subunits. Furthermore, Nayak et al.[8] recently presented an unsupervised approach to automatically learn the model for continuous basic units of signs from continuous sentences. Bowden et al. [3] try the high level description of sign language and proposed a linguistic feature vector. Wu and Huang [18] made their effort to solve the problem of view invariant posture recognition and proposed an appearance-based learning approach.

Many aspects of sign language have been studied. However, there are few attentions focusing on view invariant temporal sign language recognition. The main reason may be the difficulty of extracting view invariant features. Hindered by this difficulty, the common method of HMM[2, 7, 12, 13, 15, 16] can not be easily applied to recognize sign language in different viewpoint.

Fortunately, the difficulty of extracting view invariant features can be avoided by utilizing the epipolar geometry. The epipolar geometry studies point correspondence among different views and do not need view invariant features. It has been proved to be an efficient tool for viewpoint invariant action analysis [4, 11, 14, 19]. Following the way of utilizing the constraint of epipolar geometry, Wang et al.[17] recently proposed a novel template matching method. Their method exploits the constraint of the uniqueness of the fundamental matrix to assess if two sequences are of the same sign seen from different viewpoints. Obviously, such method requires accurate video alignment.

To relax the constraint of accurate video alignment, we propose a novel approach in this paper. The basic idea is to consider a sign as a series of tiny hand motions and utilize the HOMOGRAPHY of tiny hand motions. Using the word of "homography", we mean that there are the same tiny hand motions as well as their appearance order in different performances of the same sign. We represent a sign by a template sequence and break it up into atomic units of 3 consecutive frames. Every atomic unit record a tiny hand motion. Then for a new test sequence (taken from a potentially different viewpoint), we judge the homography by comparing every trio of frames in the sequence to each atomic unit in the template sequence. If the homography is satisfied, it is deemed that the two sequences are of the same sign. Since the basic match process is done between two 3 frames video snippets, our method does not need video alignment. Besides, our method requires no view invariant features and low training, which is same to Wang's method. However, different from Wang's method, we utilize the

different epipolar constraint to match a trio of frames with an atomic unit.

We will start by introducing the utilized epipolar constraint and step by step develop the dissimilarity measurement between an atomic unit and a trio of frames (Section 2). Based on the dissimilarity measurement, we match each atomic unit in the template sequence to every trio of frames in the test sequence. By building the match matrix and the appearance matrix, we can judge the homography for the two sequences (section 3). Next, we exploit the Nearest Neighbor rule to formalize the recognize task in Section 4. Experimental results are given in section 5. Finally, we conclude the paper in section 6.

2. MATCHING AN ATOMIC UNIT AND A TRIO OF FRAMES

In our method, an important step is to compare an atomic unit in the template sequence to a trio of frames in the input sequence. The purpose of the step is to judge whether the trio of frames and the atomic unit record the same tiny hand motion. There, we achieve this by verifying whether all pairs of feature points between the atomic unit and the trio of frames are corresponding in some stereo vision system.

To facilitate feature points extraction, we adopt color gloves. We manually label feature points (18 points in total) 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. As regard to feature point extraction from bare hands, there is a good analysis in [1, 20].

2.1. The epipolar constraint

In epipolar geometry [5, 6], there is a basis equation:

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

Where $\mathbf{p}$ and $\mathbf{p}'$ are two images of a point $\mathbf{P}$ in two views, and $\mathbf{F}$ denotes the fundamental matrix associated with the two views.

Seen from Fig. 1, there is such epipolar constraint that the set of possible matches for the point $\mathbf{p}$ is constrained to lie on the associated epipolar line l'. In other words, $\mathbf{p}'$ must lie in the line l', which also implies that

$$d(\mathbf{p}', l') = 0\,, \tag{2}$$

where $d(\mathbf{p}, l)$ denotes the spatial distance from the point $\mathbf{p}$ to the line l.

Combing Eq. 1 and Eq. 2, we can see that the line l' can be represented by $\mathbf{F}^{\mathrm{T}}\mathbf{p}$. Thus, we can rewrite Eq. 2 as

$$d(\mathbf{p}', \mathbf{F}^{\mathrm{T}}\mathbf{p}) = 0\,, \tag{3}$$

In the same manner, we have also

$$d(\mathbf{p}, \mathbf{F}\mathbf{p}') = 0\,. \tag{4}$$

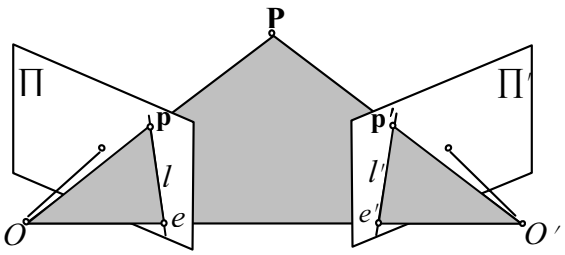

Figure 1: The epipolar geometry: the point $\mathbf{P}$, the optical centers O and O' of the two cameras, and the two images $\mathbf{p}$ and $\mathbf{p}'$ of $\mathbf{P}$ all lie in the same plane (the epipolar plane).

2.2. The dissimilarity metric

If an atomic unit and a trio of frames represent the same tiny hand motion, a fundamental matrix can be calculated by analyzing point correspondence between them. At the same time, all matched point pairs will agree with the calculated fundamental matrix. On the contrary, when an atomic unit and a trio of frames represent different tiny hand motions, the situation of the above agreement can never happen.

So we can define the dissimilarity metric between an atomic unit and a trio of frames as

$$d = \frac{1}{n}\left(\sum_{i=1}^{n}\frac{1}{2}\left[d^2(\mathbf{p}_i, \mathbf{F}\mathbf{p}'_i) + d^2(\mathbf{p}'_i, \mathbf{F}^{\mathrm{T}}\mathbf{p}_i)\right]\right). \tag{5}$$

Where n is the number of matched point pairs between the atomic unit and the trio of frames, and $\mathbf{F}$ is the calculated fundamental matrix from point correspondence between them.

Seen from the epipolar constraints in Eq. 3 and Eq. 4, d is able to reflect the degree of agreement between all matched point pairs and the calculated fundamental matrix. So we can utilize d to measure the dissimilarity between an atomic unit and a trio of frames.

3. JUDGING THE HOMOGRAPHY

As noted before, the homography means that there are the same tiny hand motions as well as their appearance order in both the template sequence and the test sequence. This section describes the detail of judging the homography.

Step 1: Build the match matrix M

We build the match matrix with row i corresponding to the i^{th} atomic unit, column j corresponding to the j^{th} trio of frames and $\mathbf{M}(i,j)$ corresponding to the matched score (Eq. 5). The i^{th} atomic unit refers to frames $(2(i\text{-}1), 2i, 2(i\text{+}1))$ in the template sequence. The j^{th} trio of frames refers to frames $(j\text{-}1, j, j\text{+}1)$ in the test sequence. Given a template sequence with m dimension and a test sequence with n dimension, we can see that the built match matrix will be $\lfloor (m-1)/2 \rfloor \times (n-2)$ dimension.

Step 2: Obtain the appearance matrix A

In fact, what we want to know is whether the tiny hand motion recorded in the j^{th} trio of frames is the possible reappearance of the tiny hand motion recorded in the i^{th} atomic unit. So we further build the appearance matrix **A** by setting $\mathbf{A}(i,j)=1$ when $\mathbf{M}(i,j)<\theta$ and setting $\mathbf{A}(i,j)=0$ when $\mathbf{M}(i,j)>\theta$, where $\theta=10.0$ is a threshold set by experiment. $\mathbf{A}(i,j)=1$ denotes that the j^{th} trio of frames is the possible reappearance of the i^{th} atomic unit.

Step 3: Search the reappearance

Furthermore, we need to know whether each atomic unit has a homographic reappearance in the test sequence. The problem is equivalent to find whether there is a monotone mapping between the set of atomic units (the template sequence) and the set of trios of frames (the test sequence), so that each atomic units and its mapped trio of frames have the same appearance order. We achieve this by designing an efficient searching algorithm as follows:

Define t as the start instant for next searching;
Define t_0 as the searched reappearance instant for the current searching;

Set t to 1;

for(each atomic unit from i=1 to $\lfloor (m-1)/2 \rfloor$)
{
  Search row i in **A** forwardly from t for the first appearance of the value 1;

  if (succeeded)
  {
    Set other elements in row i except $\mathbf{A}(i, t_0)$ to 0;
    Set other elements in column t_0 except $\mathbf{A}(i, t_0)$ to 0, so as to guarantee that one trio of frames is corresponding to only one atomic unit;
    Set t to t_0+1;
  }
  else
  {
    Search row i in **A** backwardly from t-1 for the first appearance of the value 1;
    if (succeeded)
    {
      Set other elements in row i except of $\mathbf{A}(i, t_0)$ to 0;
      Set other elements in column t_0 except of $\mathbf{A}(i, t_0)$ to 0, so as to guarantee that one trio of frames is corresponding to only one atomic unit;
    }
  }
}

After searching, the appearance matrix will give the result of reappearance for each atomic unit. $\mathbf{A}(i,j)=1$ denotes that the j^{th} trio of frames is the reappearance of the i^{th} atomic unit. The fact that all elements in row i are equal to 0 means that the i^{th} atomic unit don't appear in the test sequence.

Step 4: Judge the homography

To assess if there is a homography between the template sequence and the test sequence, we define

$$\begin{cases} P(i)=\begin{cases} j, if\ \mathbf{A}(i,j)=1; \\ -1, if\ \mathbf{A}(i,*)=0; \end{cases} \\ e_i=\sum_{\substack{k=1, \\ k\neq i}}^{\left\lfloor \frac{m-1}{2} \right\rfloor} \varepsilon_i(k),\ \varepsilon_i(k)=\begin{cases} 5, if\ P(i)=-1\ or\ P(k)=-1; \\ 0, if\ k<i\ \&\&\ P(i)<P(k) \\ \quad or\ k>i\ \&\&\ P(i)>P(k); \\ 1, if\ k<i\ \&\&\ P(i)>P(k) \\ \quad or\ k>i\ \&\&\ P(i)<P(k); \end{cases} \\ H=\frac{1}{K}\sum_{i=1,}^{K} e_i,\ K=\left\lfloor \frac{m-1}{2} \right\rfloor, \end{cases} \quad (6)$$

Seen from the above definition, $P(i)$ denotes the reappearance instant of the i^{th} atomic unit, $\varepsilon_i(k)$ means whether the k^{th} atomic unit has the earlier or later appearance order in both the template sequence and the input sequence when relative to the i^{th} atomic unit, and thus e_i contains the count information of atomic units which have different appearance orders between in the template sequence and in the input sequence when relative to the i^{th} atomic unit.

Obviously, H will be equal to 0 when there is a homography and will be greater that 0 in all other cases. Especially in the case where some atomic units don't appear in the test sequence, H will be well over 0.

4. RECOGNITION

Based on the homography metric H (Eq. 6), we exploit the nearest neighbor rule to formalize the recognition task as:

$$\begin{cases} T(S)=\arg\min_{S'\in The\ template\ set} H^{S,S'}=\arg\min_{S'\in The\ template\ set} \frac{1}{K_{S'}}\sum_{i=1}^{K_{S'}} e_i^{S,S'} \\ K_{S'}=\left\lfloor \frac{L(S')-1}{2} \right\rfloor, \end{cases} \quad (7)$$

where $L(S')$ represents the length of the template sequence S' and thus $K_{S'}$ in fact denotes the number of atomic units in S'.

5. EXPERIMENTAL RESULTS

To validate the proposed method, we test it on a 64-word vocabulary of Chinese Sign Language. Video sequences were collated for a signer performing the 64 signs. Each sign was repeated two times with one from the front view and the other from the non-front view ranging from -30° to +30°. All video sequences collated from the front view constitute the template set and the other video sequences constitute the test set. The start and end point for each sequence were labeled manually.

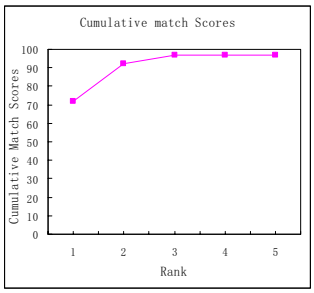

Figure 2: Cumulative match characteristic for the performance of the proposed method

For a template sequence, we firstly break it up into atomic units of 3 consecutive frames. If a template sequence has a length of m, there will be $\lfloor (m-1)/2 \rfloor$ atomic units. Then, for a new test sequence with n dimension, we build the $\lfloor (m-1)/2 \rfloor \times (n-2)$ match matrix by comparing each atomic unit in the template sequence to every trio of frames in the test sequence. Next, the related appearance matrix is obtained. By exploiting the efficient searching algorithm designed in Section 3, we can assess if the homography is satisfied and thus judge whether the two sequences are of the same sign.

To evaluate our method, we give our experimental results in terms of Cumulative Match Scores [10] in Fig. 2. From the figure, we can see that the recognition rate is 71.8% at rank 1 and 92.1% at rank 2. It demonstrates the efficiency of the proposed method.

6. CONCLUSIONS

In the paper, we have proposed a novel approach for viewpoint invariant sign language recognition. The basic idea is to see a sign as a series of tiny hand motions. We judge whether two sequences are of the same sign by assessing if there are the homographic tiny hand motions between them. The proposed approach requires no view invariant features, low training and no alignment.

The future work includes testing the performance of the proposed approach on a larger sign language vocabulary, detecting automatically the start and end point for each sign and investigating the possibility of considering each tiny hand motion as a singme and integrating with statistic methods.

ACKNOWLEDGES

This research is supported by the National Science Council, R.O.China, under the Grant 60533030. The authors would like to thank Miss Yan Ma, who provides great help on data collection.

7. REFERENCES

[1] Antonis A. Argyros, and Manolis I.A. Lourakis, "Real-Time Tracking of Multiple Skin-Colored Objects with a Possibly Moving Camera," In Proceedings of The 8th European Conference on Computer Vision, pp. 368-379, 2004.

[2] B. Bauer, and K.F. Kraiss, "Video-Based Sign Recognition Using Self-Organizing Subunits," International Conference on Pattern Recognition, pp. 434-437, 2002.

[3] R. Bowden, D. Windridge, and et al., "A Linguistic Feature vector for the Visual Interpretation of Sign Language", European Conference on Computer Vision, pp. 390-401, 2004.

[4] A. Gritai, Y. Sheikh, and M. Shah. "On the use of Anthropometry in the Invariant Analysis of Human Actions", Int. Conf. on Pattern Recognition, pp. 923-926, 2004.

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

[6] Q. T. Luong, and O. D. Faugeras, "Self Calibration of a Moving Camera from Point Correspondences and Fundamental Matrices," in IJCV, 22(3), pp. 261-289, 1997.

[7] R.M. McGuire, T. Starner, and et al.. "Towards a one-way American sign language translator," International Conference on Automatic Face and Gesture Recognition, pp.620 – 625, 2004.

[8] S. Nayak, S. Sarkar, and Barbara Loeding, "Unsupervised Modeling of Signs Embedded in Continuous Sentences," IEEE Conf. Computer Vision and Pattern Recognition, 2005.

[9] 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, 27(6): 873-891, 2005.

[10] J. Phillips, H.Moon, S. Rizvi, and P. Rause, "The FERET evaluation methodology for face recognition algorithms," IEEE Trans. Pattern Analysis and Machine Intelligence, 22: 1090-1104, 2000.

[11] C. Rao, A. Gritai, M. Shah, and T. Syeda-Mahmood, "View invariant alignment and matching of video sequences," International Conference on Computer Vision, pp. 939-945, 2003.

[12] T. Starner, and A. Pentland, "Visual recognition of american sign language using hidden markov models," International Conference on Automatic Face and Gesture Recognition, pp. 189-194, 1995.

[13] T. Starner, J. Weaver, and A. Pentland, "Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video", IEEE Trans. Pattern Analysis and Machine Intelligence, 20(12): 1371-1375, 1998.

[14] T. Syeda-Mahmood, A. Vasilescu, and S. Sethi, "Recognizing action events from multiple viewpoints," IEEE Workshop on Detection and Recognition of Events in Video, pp. 2001.

[15] C. Vogler, and D. Metaxas, "Asl recognition based on a coupling between hmms and 3d motion analysis," International Conference on Computer Vision, pp. 363–369, 1998.

[16] C. Vogler, and D. Metaxas, "A Framework for Recognizing the Simultaneous Aspects of American Sign Language," Computer Vision Image Understanding, 81: 358-384, 2001.

[17] Q. Wang, X. Chen, L. Zhang, C. Wang, and W. Gao. "Viewpoint invariant sign language recognition," International Conference on Image Processing, 2005.

[18] Y. Wu, and T. S. Huang, "View-independent recognition of hand postures," IEEE Conf. Computer Vision and Pattern Recognition, pp. 88-94, 2000.

[19] A. Yilmaz, and M. Shah, "Recognizing Human Actions in Videos Acquired by Uncalibrated Moving Cameras," International Conference on Computer Vision, 2005.

[20] X. Yin, and M. Xie, "Estimation of the fundamental matrix from uncalibrated stereo hand images for 3D hand gesture recognition," Pattern Recognition: 36(3), pp. 567-584(18) , 2003.