15—2

MVA '98 IAPR Workshop on Machine Vision Applications, Nov. 17-19, 1998, Makuhari, Chiba, Japan

Vision-based Interpretation of Hand Gestures by Modeling Appearance Changes in Image Sequences

Yuanxin Zhu, Yu Huang, Guangyou Xu, Haibing Ren, Zhen Wen
Department of Computer Science and Technology, Tsinghua University
Email: {zyx,yh}@vision.cs.tsinghua.edu.cn, xgy-dcs@mail.tsinghua.edu.cn

Abstract

Vision-based interpretation of hand gesture had typically made use of high-level parametric models representing the body parts such as arms, figures, palms etc. and their connections to each other. Such 3-D model-based interpretation has been successful in some case; however, heavy computation makes this approach a very difficult task. Recently appearance-based methods are more widely used for their computational efficiency. By use of variable-order parameterized models of image motion and robust dominant motion regression, in this paper, we propose a motion-based segmentation scheme to feature extraction of hand gestures. The proposed scheme can directly estimate image motion of an object in two unsegmented images and obtain fine segmentation of that object from background at the same time. Based on inter-frame image motion parameters and the fine segmentation, we can construct various motion features, shape features, or their combinations for the purpose of hand gesture interpretation. With these features 12 kind of hand gestures used in our experiment can be reliably interpreted.

1 Introduction

Although automatic speech recognition has been a topic of research for decades, it has only been in recent years that there has been an increased interest in visual interpretation of hand gestures [1]. The key motivation for visual interpretation of hand gestures is to introduce this natural and intuitive communication mode to human computer interface (HCI). Besides that, visual interpretation of hand gestures can be applied to virtual reality, 3-D design, telepresence, interaction and visualization, medical research etc. Approaches to visual interpretation of hand gestures can also be directly extended to interpretation of facial expression, lip-reading, body gesture interpretation, spatial-temporal texture classification, image mosaics and content-based image database retrieval.

Approaches to interpretation of hand gestures fall into two major groups: those based on 3-D hand/arm modeling and those based on appearance change modeling. The 3-D hand and arm models have often been a choice for hand gesture modeling [2, 3], but heavy computation make these approaches a very difficult task and far away from application. Appearance-based approaches mainly address how to recognize hand gestures directly from visual appearance changes in image sequences. Their focal points are the appearance changes caused by image motion rather than the static 3-D structures of hand/arm; therefore they are computationally efficient and more widely used [4]. Darrell uses a direct approach which represents the object performing gestures with a vector of similarity scores to a set of 2D spatial views (or 3D spatio-temporal views), but it cannot be expected to perform generalization across multiple users or when a single user performs a gesture that is not well modeled by previously-seen spatial-temporal patterns [5]. Quek computed the hand motion path (no recognition results), yet shape or spatial information was lost [6]. Cui et. al. proposed a framework that addresses three key aspects of the hand sign interpretation, that is the hand shape, the location, and the movement. The framework has been tested to recognize 28 different hand signs [7]. In addition, Black addresses the multiple affine motion estimations by robust dominant motion regression in [8] and optical flow estimation in segmented images using variable-order parameterized models with local deformations in [9].

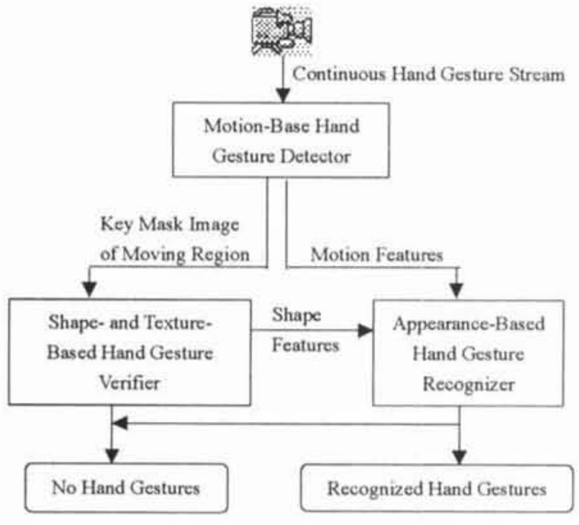
Continuous Hand Gesture Stream
Motion-Base Hand Gesture Detector
Key Mask Image of Moving Region
Motion Features
Shape- and Texture-Based Hand Gesture Verifier
Shape Features
Appearance-Based Hand Gesture Recognizer
No Hand Gestures
Recognized Hand Gestures

Figure 1. Processing architecture of on-line hand gesture interpretation system

Our approach belongs to the second group. Compared with previous work, first, we propose a new method for motion estimation in two unsegmented images using variable-order parameterized models of image motion and robust regression. Second, we show how the estimated motion parameters can be combined to model appearance changes in image sequences. Third, we propose to integrate motion, shape, and texture information to perform segmentation and interpretation of hand gestures.

Finally, we demonstrate an on-line system for interpretation of twelve hand gestures.

Fig.1 presents an overview of the processing architecture of our on-line hand gesture interpretation system. Base on the appearance change in an image sequence, the first step roughly detects whether there is a predefined hang gesture contained in the current sequence. If the first step determined that there exists one hand gesture, then the second step begins to work immediately. It performs hand gesture verification by shape and texture analysis of the key mask image of the gesturing hand generated by the first step. If the first step detects and the second step verifies that a hand gesture does appear in the image sequence, then the motion and shape features, generated by the two steps, will be delivered to the third step (hand gesture recognizer) to finally determine the type of the gesture.

In the following section, we describe motion estimation in two unsegmented images. Modeling appearance change in image sequences is given in Section 3. Hand gesture segmentation and interpretation algorithms are explained in Section 4. Section 5 gives an example of motion estimation and experimental results are arranged in Section 6. The article is concluded in Section 7.

2 Motion Estimation in Unsegmented Images

Parameterized Modes of Image Motion

The well-known parameterized models of image motion are the translation model, the affine model and the planar model [10]. For example, the affine model can be written as

$$\mathbf{u}(\mathbf{x}) = \begin{bmatrix} u(x,y) \\ v(x,y) \end{bmatrix} = \begin{bmatrix} a_0 + a_1 x + a_2 y \\ a_3 + a_4 x + a_5 y \end{bmatrix}$$

Where the a_i are constants and where $u(x,y)$ and $v(x,y)$ are the horizontal and vertical components of the flow at the image point $\mathbf{x} = (x,y)$. For convenience of notation we define

$$\mathbf{X}(\mathbf{x}) = \begin{bmatrix} 1 & x & y & 0 & 0 & 0 & x^2 & xy \\ 0 & 0 & 0 & 1 & x & y & xy & y^2 \end{bmatrix},$$

$$\mathbf{T} = (a_0, 0, 0, a_3, 0, 0, 0, 0)^T,$$

$$\mathbf{A} = (a_0, a_1, a_2, a_3, a_4, a_5, 0, 0)^T,$$

$$\mathbf{P} = (a_0, a_1, a_2, a_3, a_4, a_5, \mathrm{a}_6, \mathrm{a}_7)^T.$$

such that $\mathbf{u(x,T)} = \mathbf{X(x)T}$, $\mathbf{u(x,A)} = \mathbf{X(x)A}$ and $\mathbf{u(x,P)} = \mathbf{X(x)P}$ represent, respectively, the translation, the affine and the planar models described above. Given the model of image motion we employ robust regression scheme [8] to recover those parameters. To cope with large motions a coarse-to-fine strategy [10] is used.

Motion Estimation by Twice Robust Regressions

As identified by [11], the main goal of robust statistics is to recover the structure (in our case the parameterized models of image motion) that best fits the majority of the data while identifying "inliers" and rejecting "outliers" or "deviating substructures".

To do motion estimation in two unsegmented images, we explore the use of variable-order parameterized models of image motion and robust dominant motion regression. The proposed method consists of two steps (for details see [12]). In the first step, our goal is obtain a coarse segmentation of the moving object from complex background. Since we assume the background maintains static roughly, we perform the first robust regression with the full image as the region of analysis and with choosing the translation model. The regressed dominant motion should be (actually is) the motion of the background since most pixels belong to the background. Once the dominant motion has been determined we can obtain a coarse segmentation (including some noise points) of the moving object from the background by analyzing the residual, that is by identifying outliers according to the result of the translation regression.

The goal of the second step is to achieve accurate motion estimation and fine segmentation of the object at the same time. We use a 2-D affine model or planar model of image motion to approximate the projected 3-D motions of the object onto the image plane. This assumption is valid when the difference in depth caused by the motion is small relative to the distance of the object from the camera. With the coarse segmentation of the moving object as our new analysis region and with the affine or planar model, we perform the second robust regression. Since almost all pixels belong to the moving object at this time the regressed motion is the motion of that object. At the same time, we can achieve a fine segmentation of the moving object by identifying inliers and rejecting outliers according to the result of the second regression. Therefore, we can attain accurate motion estimation in unsegmented images by using variable-order-parameterized models of image motion and robust regression.

3 Modeling Appearance Changes in Image Sequences

Since we consider only hand gestures performed purposefully by the user, gestures must be deliberately aimed at the machine's vision system. This eliminates most of the problems of three-dimensional occlusion and disocclusion from the modeling process. Given an image sequence containing a hand gesture, Let T represents the temporal length of the sequence (number of frames in the sequence) and let $I_t\ (t = 1,2...,\mathrm{T})$ be the tth frame image. Let $\mathbf{m}[t]$ be the motion vector between frame I_t and frame I_{t+1}. As described in [13], the parameters a_i of image motion models have simple interpretations in terms of image motion. We can express horizontal translation (u), vertical translation (v), isotropic expansion (e), deformation (d), rotation about the viewing direction (r), yaw (y), and pitch (p) as combinations of the a_i :

$$u = a_0;\ v = a_3;\ e = a_1 + a_5;\ d = a_1 - a_5$$

$$r = -a_2 + a_4;\ y = a_6;\ p = a_7.$$

So we can model the inter-frame appearance change by defining a motion vector $\mathbf{m}[t]$ as $[u, v, e, d, r, y, p]^T$.

We assert that there exists a *consistent appearance change* in the sequence if the angles between two consecutive motion vectors $\mathbf{m}[t]$ and $\mathbf{m}[t+1]$ are less than a predefined threshold θ_0. In order to eliminate variation of gesturing rate in the sequence, we model the appearance change in the sequence as a weighted and normalized summation over all these inter-frame appearance changes, that is,

$$\Sigma = \hat{\mathbf{m}} \bullet \sum_{t=1}^{\mathrm{T}-1} \mathbf{m}[t], \qquad \mathbf{M} = \frac{\Sigma}{\|\Sigma\|}$$

where $\mathbf{M}$ is the appearance change in the sequence (we name it overall image motion representation), $\hat{\mathbf{m}}$ is a weight vector, and "$\bullet$" represents inner product operation.

4 Gesture Segmentation and Interpretation

Continuous hand gesture interpretation is much harder than isolated hand gesture recognition. There is no silence between the hand gestures, so segmentation of hand gesture streams is an unsolved difficult task. In this paper, we propose a two-step segmentation approach to address this problem partially. The proposed two-step segmentation approach consists of motion-based detection of hand gestures and shape- and texture- based verification, which looks much like a prediction-verification procedure. In the first step, we assume that an image sequence contains a hand gesture if the sequence has the following four characteristics.

1. There is only one moving object in the image sequence.
2. The moving object doesn't move from surround toward center of the image.
3. There exist a consistent appearance change in the sequence;
4. The temporal length of the sequence is longer than T_1 but not more than T_2.

As described above, We can segment out the moving region according to the result of motion estimation. After eliminating isolated points and filling in holes in the moving region, a mask image of the moving object from every two consecutive frames can be created.

Suppose an image sequence containing a hand gesture has been detected by the first step, the second step begins to work immediately. Based on the number of points contained in a mask image the key mask image that best represents the hand shape of the corresponding gesture is selected from the mask image sequence. Then we perform verification of hand gestures by shape and texture analysis of the key mask image.

Given a training set of isolated hand gestures, we create an appearance change reference template for each hand gesture by a minimax type of optimization based on the Euclidean distance [14]. Then template-based classification technique is employed to perform interpretation of hand gestures.

5 Motion Estimation Example

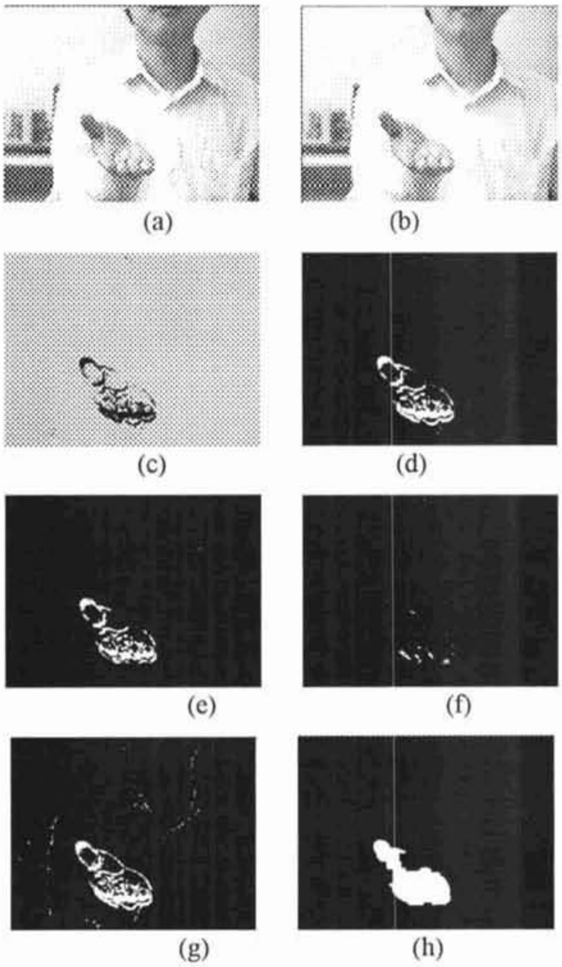
(a) (b) (c) (d) (e) (f) (g) (h)

Figure 2: An example of motion estimation.

A motion estimation example is shown in Fig. 2 where

(a) and (b) are the 2nd and 3rd frames in an image sequence containing an "move up" hand gesture. The inliers identified according to the result of the first translation motion regression is shown as gray pixel and the outliers as dark pixels in (c). The region consisting of white pixels shown in (d) is the coarse segmentation of the gesturing hand. The parameters computed by the second regression with the affine model are {-0.3644, 0.0040, 0.0066, -2.2835, -0.0012, -0.0086, 0, 0}, So the motion vector between the two images is {-0.3644, -2.2835, -0.3730, -0.0052, -0.3558, 0, 0}. The fine segmentation of the gesturing hand (inliers identified according to the result of the second regression) is shown as white pixels in (e), and outliers in (f). The image shown in (g) is a thresholded difference image between (a) and (b). The image shown in (h) is the key mask image of the hand gesture. Compared (e) with (g), it is obvious that there are many noise pixels in (g) but few in (e).

6 Experimental Results

We conceive to use hand gestures as a 3-D pointing device for more natural and intuitive HCI. So the twelve hand gestures used in our demonstration system are "move up", "move down", move left", "move right", "move farther", "move nearer", "yaw left", "yaw right", "curl clockwise", "curl counterclockwise", "pitch down", and "pitch up".

We carried out a large set of experiments to verify and evaluate the effectiveness of the overall image motion representations and the recognition procedure proposed in this paper. We ask ten subjects to perform the twelve hand gestures in front of a digital-camera. From these subjects we collected a training set of 120 image sequences. Each sequence is about one second long and contains one hand gesture. Images are 160×120 pixels with 256 gray-level (taken at 10Hz). Time for recognition of one hand gesture is 5s or so on a PII (266Hz) PC. Average recognition rate achieved on the training set is over 92%. Average recognition rate attained on on-line hand gesture interpretation system is over 85%.

7 Conclusion

Based on variable-order parameterized models of image motion and robust regression, in this paper we presented a direct method of motion estimation in unsegmented images and show how to integrate motion, shape, and texture information to perform segmentation and interpretation of hand gestures. Our on-line system demonstrates the effectiveness of the proposed methods and show that the twelve hand gestures can be recognized with high accuracy based on the proposed methods. Choice of variable-order parameterized models of image motion enables efficient motion computations and is numerically stable. In addition, with robust regression erroneous measurements at motion boundaries are treated as outliers and their influence is reduced. To improve recognition rate further, we plan to define more descriminative shape features to create more descriptive appearance change models in the future.

References

[1] ZHU Yuanxin, XU Guangyou, Chih-Ho Yu, Vision-based hand gesture interpretation: a survey, Proceedings of 3rd China Conference on Artificial Interface and Artificial Application, pp. 279-284, Aug. 1997 (in Chinese).

[2] T. F. Cootes, C. J. Taylor, and J. Graham, "Active shape models--their training and application," Computer Vision and Image Understanding, Vol.61, pp.38-59, Jan. 1995.

[3] T. Heap and D. Hogg, "Towards 3D hand tracking using a deformable model," Int'l Conf. Automatic Face and Gesture Interpretation, Killington, Vt., pp.140-145, Oct. 1996.

[4] Pavlovic, V., Sharma, R., and Huang, T. S., Visual interpretation of hand gestures for human-computer interaction: a review, IEEE Trans. PAMI, Vol. 19, No. 7, July, 1997.

[5] T. J. Darrell, I. A. Essa, A. P. Pentland, "Task-specific gesture analysis in real-time using interpolated views," IEEE Trans. PAMI, Vol.16, No.12, 1996.

[6] Quek F., Unencumbered Gestural Interaction, IEEE Multimedia, 1996: 36-47.

[7] Cui Y. and Weng J. J., Hand sign recognition from intensity image sequences with complex background, Intl. Conf. on Automatic Face and Gesture Recognition, Killington, Vt., pp.259-264, Oct. 1996.

[8] Black, M. J., Anandan, P., The robust estimation of multiple motions: parametric and piecewise-smooth flow fields, Computer Vision and Image Understanding, 63 (1): 75-104, 1996.

[9] Black M J, Jepson A D. Estimation optical flow in segmented images using variable-order parametric models with local deformation. IEEE Trans. PAMI, 1996, 18(10): 972-986.

[10] Bergen, J. R., Keith, P., Hanna, J., and Hingorani, R., Hierarchical model-based motion estimation, ECCV'92, pp. 237-252, 1992.

[11] F. R. Hampel, et. al., Robust Statistics: The Approaches Based On Influence Functions, Wiley, New York, 1986.

[12] ZHU Yuanxin, HUANG Yu, XU Guangyou, Chih-Ho Yu, "Motion-based segmentation scheme to feature extraction of hand gestures", Proc. SPIE vol. 3545, ISMIP'98, Wuhan, China 1998, To appear.

[13] Black, M. J., Yacoob, Y., Tracking and recognizing rigid and non-rigid face motion using local parametric models of image motion, ICCV'95, pp. 374-381, 1995.

[14] Rabiner, L. R., On Creating reference templates for speaker independent interpretation of isolated words, IEEE Trans. ASSP, Vol. 26, No. 1, February 1978.