Motion Estimation Based on Optical Flow and ANN

Jiafeng. Zhang, *Feifei. Zhang, *Masanori. Ito

Tokyo University of Marine Science and Technology, 2-1-6, Etchujima Koto-Ku, Tokyo
(Tel : 81-3-5245-7422; Fax : 81-3-5245-7422)
(fnfnzhang78@yahoo.co.jp)

Abstract: Motion estimation provides an attractive choice to cumbersome interface devices for human computer interaction (HCI). Worthy of note, visual recognition of hand gestures can help achieving in an easy and natural way for interaction between human and computer. The interfaces of HCI and other virtual reality systems depend on accurate, real-time hand and fingertip tracking for association between real objects and relative digital information. They cost expensively and complicated operations make them troublesome. We are developing a real-time, view-based gesture recognition system. Optical flow is estimated and segmented into motion fragments. Using artificial neural network (ANN), it can compute and estimate the motion of gesture. Comparing with the traditional approaches, theoretical and experimental results show that this method has more simple hardware and algorithm demands but more effectiveness. It can be used in moving object recognition system for understanding the human body languages.

Keywords: motion estimation; optical flow; ANN**.**

I. INTRODUCTION

In order to meet human requirements better, motion estimation at a distance has recently gained more and more interests from computer vision researchers. It has a particularly attractive modality from a surveillance perspective, such as human computer interaction (HCI) or more generally human-machine interaction (HMI) etc.

Up to now, the most popular modes of HCI/HMI are based on mechanical devices, such as motion-capture system, remote controller or fingertip-tracking system. All these devices have shown more and more disadvantages such as inconvenience, high expense, complex manipulation, and even causing injuries to users.

To make it more humanized, a natural approach of speech recognition has been employed in HCI/HMI [1]. Although the vocabulary used in a speech HCI/HMI is much less than in human-human interaction (HHI), it is too hard to make a speech HCI/HMI applicable for every linguistic form of variety of languages, even for individual accents or localism. And, to make use of speech recognizing of HCI/HMI is not a good choice in a place where silence is required, or a noisy surrounding. Relevant experiments have shown that in a noisy surrounding blind people can only understand 23% of the speech contents averagely, while the others who are not blinded can understand 65% with the same speech. It proves that visual and auditory sensations are both important in HHI, and human vision system does great help for speech understanding. In fact, body motion and gesture are kinds of shape languages which do not depend on the individua1 characters or various linguistic forms, on which voice languages do. To make the communication between human and machine more natural and humanized, a lot of works have been spent to motion estimation and gesture recognition of HCI/HMI.

General motions include walking, jumping, turning around or gestures such as moving up, moving down, waving hand etc. can be divided into a series of motion states corresponding a image sequence. According to these states, human can easily track and recognize motions or gestures. How do we make computer be able to replicate such recognition ability? Initially motion analysis was devoted to the complete recovery of motion information from image sequences known as an ill-posed problem [2].

However, it is not necessary to obtain information so detailed for further analysis of dynamic content in image sequences. On the other hand, in terms of various motions, necessary information for the motion may usually change. In this paper, a visual approach based on gesture recognition is proposed.

To recognize a hand gesture from a complex backside image essentially, the three stages of detection, segmentation and recognition for a number of gestures which are assigned as some control commands must be included (Fig.1).

Details of the method and preliminary experimental results will be shown below.

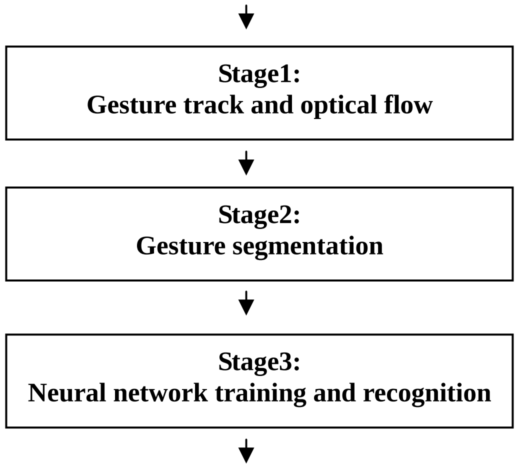


Fig.1. Three stages of gesture recognition

II. OPTICAL FLOW ESTIMATION

The major milestone of the concept of optical flow was firstly advanced and attempted to calculate by Horn and Schunck. They define optical flow as "the distribution of apparent velocities of movement of brightness patterns in an image" [3]. If the velocities of the brightness patterns (objects within the image) are known, then a robot or vision system using the optical flow techniques will have some knowledge of how its surroundings are changing.

To compute the optical flow, we assume that the object being imaged is a flat surface and the illumination on the image is constant and uniform. It is also assumed that the reflectance of the object varies smoothly and has no spatial discontinuities. These assumptions assure that the image brightness or intensity is differentiable.

Let $I = I(x, y, t)$ denote the continuous space-time intensity distribution. If the intensity remains constant along a motion path, that is,

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \qquad (1)$$

then $dI(x, y, t)/dt = 0$. This latter condition can also be written as

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0 \qquad (2)$$

where $u = \dfrac{dx}{dt}$ and $v = \dfrac{dy}{dt}$ denote the components of the coordinate velocity vector in terms of the continuous spatial coordinates. Equation (2) is known as the optical flow equation (OFE). The OFE is not sufficient to uniquely specify the 2-D velocity field. The remainder of this section outlines the methods employed in this study to estimate the velocity field.

1. Horn and Schunck method

The Horn and Schunck method seeks a motion field that satisfies the OFE with the minimum pixel-to-pixel variation among the velocity vectors. The pixel-to-pixel variation of the velocity vectors can be quantified by

$$\iint [(I_x u + I_y v + I_t)^2 + \lambda^2 (\|\nabla_u\|_2^2 + \|\nabla_v\|_2^2)] dxdy \qquad (3)$$

Horn and Shunck proposed the following iteration to estimate the optical flow

$$u_{n+1} = u_n - \frac{I_x[I_x u_n + I_y v_n + I_t]}{\alpha^2 + I_x^2 + I_y^2}$$
$$v_{n+1} = v_n - \frac{I_y[I_x u_n + I_y v_n + I_t]}{\alpha^2 + I_x^2 + I_y^2} \qquad (4)$$

where n is the iteration counter and all partial derivatives are evaluated at the point (x, y, t)

2. Lucas and Kanade method

Following Lucas and Kanade [4], we implemented a weighted least-squares (LS) fit of local first-order constraints (2) to a constant model for (u, v) in each small spatial neighborhood D by minimizing

$$\sum_D W^2(x, y)(I_x u + I_y v + I_t)^2 \qquad (5)$$

where $W(x, y)$ denotes a window function that gives more influence to constraints at the center of the neighborhood than those at the periphery. I_x, I_y, I_t denote the partial derivative of $I(x, y, t)$ at point (x_i, y_i). Using LS, the solution to (5) is given by

$$u\sum_D W^2(x, y)I_x^2 + v\sum_D W^2(x, y)I_x I_y + \sum_D W^2(x, y)I_t I_x = 0 \qquad (6)$$

$$u\sum_D W^2(x, y)I_x I_y + v\sum_D W^2(x, y)I_y^2 + \sum_D W^2(x, y)I_t I_y = 0 \qquad (7)$$

The solution to (6) and (7) is

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \sum W^2(x, y)I_x^2 & \sum W^2(x, y)I_x I_y \\ \sum W^2(x, y)I_x I_y & \sum W^2(x, y)I_y^2 \end{pmatrix}^{-1} \begin{pmatrix} -\sum W^2(x, y)I_t \quad I_x \\ -\sum W^2(x, y)I_t \quad I_y \end{pmatrix} \qquad (8)$$

Barron et al[5] compared many computing approach used for optical flow calculation and Lucas and Kanade method shows better veracity and stability. Further more, it is relatively speedy and easy in computation and implementation, and we adopt it in this paper.

3. Real-time gesture tracking and optical flow

In our experiment system, a standard 2.4GHz Pentium 4 PC was used with a monochrome Sony XC-HR50 digital video camera.

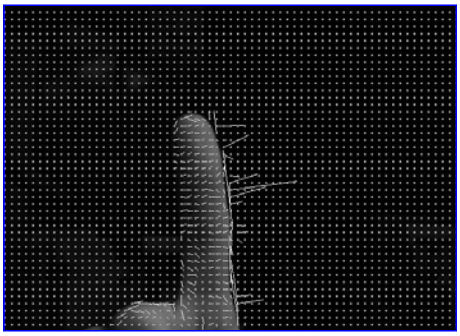


Fig.2. The optical flow of a forefinger motion (from left to right)

Fig.2 shows the moving velocity (from left to right) extracted from the real-time video stream by utilizing the Lucas and Kanade optical flow method outlined in this paper. The values of velocity are indicated by lines (the rectangles denote the origin in the former frame). Fig.2 also shows the velocity field of moving area.

4. Flow estimation confidence measurement

There exists a significant problem outlined in sections II.1-II.2, that is, for those methods that integrate normal constraints with global (regularization) smoothness constraints, these measurements are lack of confidence.

In Fig.2, we can see undesired optical flow on the edges of the moving parts. This kind of undesired optical flow is evident in the process using matching methods or with some degree of dilation and so on.

To evaluate the confidence in the flow estimation, for example, Anandan [6] fitted a cubic surface to the SSD values and defined a confidence measurement based on the curvatures of the surface. Another way to verify the optical flow was to use the left-right check [7], which is, computing the flow for frame I_t with respect to I_{t+1}, then repeating for I_{t+1} with respect to I_t. Only the flow whose estimation results were same (or close) should be used. Although these techniques gave good results, they required massive calculations. Here, we use a simple way depending on the histogram of optical flow to eliminate the common flow estimate errors, and it has been proved that for our later analysis this technique has given sufficient flow estimation.

Fig.3 shows the flow histogram in x-direction and y-direction; and Fig.4 only shows the most important optical flow we expected by removing the flow noise in optical flow field through histogram.

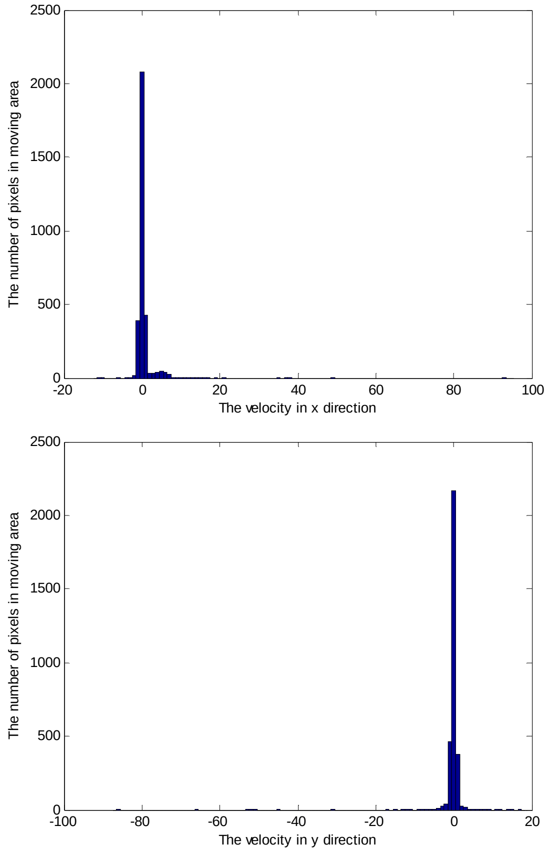


Fig.3. The histogram of flow velocity

Fig.4. Using histogram to get the main flow

From the histogram of Fig.3, we can set a threshold to get rid of the undesirable flow.

III. SEGMENTATION AND RECOGNITION

1. Gesture segmentation with optical flow

The first task is to work out how the optical flow vectors we get from each frame in II can be presented to the ANN. In our experiments, we represented a kind of motion segmentation as a series of 12 vectors, which come from consecutive video frames. Fig.5 shows how the gesture for Right Arrow can be represented as a series of vectors.

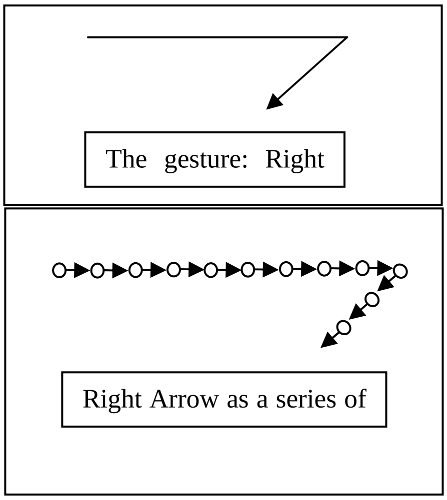
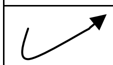
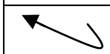
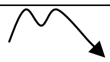
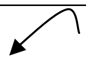
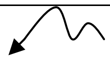
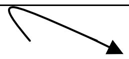
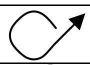
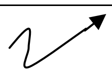
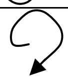


Fig.5. Gestures as vectors

Each vector can be got by calculating the mean value of optical flow in the corresponding frame. To aid training, these vectors are normalized before becoming part of the training set. Therefore, all the inputs into the network have been standardized. In this way, additional advantage is that, when we come to process the gestures made by users, we can achieve even distribution of the vectors through the gesture pattern, which will aid the ANN in the recognition process.

In our experiment, we predefined 10 gestures to express different semantic.

Table1. Predefined 10 basic gestures for semantic

Gesture	Command	Gesture	Command
	Begin		Down
	Stop		Right
	Forward		Left
	Backward		Take up
	Up		Put down

IV. Recognition by BP network

Backpropagation (BP network)[8], works as follows: First create a network with one or more hidden layers and randomize all the weights. Then present a pattern to the network and compute error value. This error value is then used to determine how the weights from the layer below the output layer are adjusted. Once the weights for the current layer have been adjusted, the same thing is repeated for the previous layer and so on until the first hidden layer is reached. The next time the input pattern is presented, the output will be a little bit closer to the target output. This whole process is then repeated with all the different input patterns many times until the error value is coming to an acceptable limit for the recognition system.

Using BP network, the recognition rate achieved 83.3% in our experiments for the predefined gestures.

V. CONCLUSION

In this paper, we have utilized optical flow method to segment human's gesture motions from real-time video stream. A histogram technique for getting rid of flow noise is proposed, and using BP networks, we can recognize a set of gestures, which are defined by optical flow vectors in a context of motion segmentation. Using this system, the recognition rate achieved 83.3% in our experiments for the predefined gestures. Future work for this system includes: adding more gestures and enhancing recognition rate; using HMM to make the gesture recognition more robust.

REFERENCES

[1] Pavlovi'c VI, Sharma R, Huang TS (1997) Visual interpretation of hand gestures for human-computer interaction: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence 19:677-695
[2] Bertero M, Poggio T, Torre V (1988) Ill-posed problems in early vision. Proc. of the IEEE 76(8): 869-890
[3] Horn BKP, Schunk BG (1981) Determining optical flow. Artificial Intelligence 17:185–203
[4] Lucas B, Kanade T (1981) An iterative image registration technique with an application to stereo vision. DARPA Image Understanding Workshop 2:121--130
[5] Barron JL, Fleet DJ, Beauchemin SS, Burkitt TA (1992) Performance of optical flow techniques. CVPR
[6] Anandan P (1989) A computational framework and an algorithm for the measurement of visual motion. International Journal of Computer Vision 2:283–310
[7] Faugeras O, Hotz B, Mathieu H et al. (1993) Real time correlation based stereo: algorithm implementations and applications. Technical Report 2013
[8] Freeman JA, Skapura DM (1992) Backpropagation Neural Networks Algorithm Applications and Programming Techniques, pp 89-125.