Multi-dimensional game interface with stereo vision

Yufeng Chen, Mandun Zhang, Peng Lu, Xiangyong Zeng, and Yangsheng Wang

Institute of Automation, Chinese Academy of Sciences**
100080 Beijing, P.R.China
yufeng.chen@ia.ac.cn {mzhang,plu,xzeng,wys}@mail.pattek.com.cn

Abstract. An novel stereo vision tracking method is proposed to implement an interactive Human Computer Interface(HCI). Firstly, a feature detection method is introduced to accurately obtain the location and orientation of the feature in an efficient way. Secondly, a searching method is carried out, which uses probability in the time, frequency or color space to optimize the searching strategy. Then the 3D information is retrieved by the calibration and triangulation process. Up to 5 degrees of freedom(DOFs) can be achieved from a single feature, compared with the other methods, including the coordinates in 3D space and the orientation information. Experiments show that the method is efficient and robust for the real time game interface.

1 Introduction

Computer vision is a rapid developing area with more and more application requirements. One of the most basic functions is to interact with human. Taking advantage of convenience and natural interview compared with other advanced technology such as mechanical or electromagnetic, they are already implemented on some advanced HCI for special purpose to take the place of or aid the traditional devices such as mouse. Specially, their interactive capabilities are more suitable to be fully used on games.

1.1 Previous works

Many works has been reported to be used in the related areas. Some typical early works has been introduced by Gavrila and Freeman [1, 2].

One of the most important steps of vision interaction is tracking and detection. Some body features are tracked as camera mouse[3] to help people with severe disabilities. And recently, eyekeys[4] are used to detect the gaze in real-time. Many other features are tried such as face, gesture[5], and even the body[8].

It has been evaluated[6] that salient features are needed to get better location accuracy, on the other hand these features are hard to track the irregular human movement while need more computing to search in a large area. Some correlation

** Funded by the National 863 project.

methods[3] are recommended to use under this situation, but the speed is another obstacle to the method.

The first part of our work is to propose a novel local statistic method to get an efficient location of the salient features. Combined with multiple cues, such as motion, color and intensity features, the searching amount are largely reduced and the feature can be precisely located.

Some related works had been carried out by Zhang[16] and Wu[10], they use hand as a simple input device. Although many games are using vision based techniques such as Eyetoy, the use of human movement information is limited.

To get more information of the hand movement, more cameras are needed to get 3D information. 3D interactive is reported [9] with a graphic point of view, which shows the requirement of the nature multi-dimensional method. Many stereo vision based applications are concentrating on the large features, such as human figure detection[11], to solve the occlusion problems.

Our efforts aimed to get more dimensional information from simple features. Up to five DOFs can be achieved from the stereo vision. This is very useful to control the complex object in the game.

The novel feature detection method is introduced in the section 2, an optimized searching method is proposed in the section 3, and the stereo vision with 5 DOFs is proposed in the section 4. Then the experiments are carried out to show the efficiency and accuracy of the method. At last the paper ends with a conclusion and some prospects.

2 Feature detection

As discussed above, some properties are required for interactive games: Firstly the invariant should be kept under different conditions such as transfer, rotation and lighting. Secondly, it should be sensitive to the feature difference which helps to improve the accurate of the feature location[6]. The efficiency is also required for the real-time game interface.

In this paper we introduce a local moment based feature detection method. The invariant moments is first introduced as Hu Moments[12], many other moments[13, 14] are proposed to improve the performance.

However, all the moments mentioned above must treat the object as a whole, which means the model should be rigid and well segmented or featured without occlusion, this requirement is not always met. Takamatsu[15] tried to use local moment to recognize the parts of the object as an improvement, but more work related with local properties is still lacking.

Given the perspective transformation, the light amount from a certain view point depends on three main factors: the normal of the surface N, the lighting condition L and the albedo a. So that the image can be derived

$$I(x,y) = C(a(x,y)N(x,y)L(x,y))$$

Where Function $C(\cdot)$ is the integration transform of the camera sensor, which is always considered as linear integration if the exposure time is suitable.

From the model of the image above, we can see clearly that the image, which depend on the lighting condition and camera system, is not uniquely determined by the feature. It is described by the surface normal N and the albedo a. Fortunately the camera system can be simplified as a linear integration, and the light condition could be viewed as a combination of many point light sources, which are also linear both in amount and their distribution. Thus the feature invariant is formed as follows:

Given a standard feature$\hat{f}(x, y)$ depending only on the object properties, a real feature image supposed to be effected by a local multiplicative transformation, which is corresponded with the different contrast, and a linear spaced additive transform, which is introduced by the lighting conditions.

$$f(x, y) = k \times \hat{f}(x, y) + a + bx + cy$$

Here we use the common Cartesian moment for simple explanation of the method.

$$m_{p,q} = \int_{x_0}^{x_d} \int_{y_0}^{y_d} x^p y^q f(x, y) dx dy$$

Where the $f(x, y)$ is the two dimensional function and x_0, x_d, y_0, y_d are the border of the target window to intergraded, and p, q stand for the moment orders.

If the equation below is met,

$$x_0 + x_d = 0, y_0 + y_d = 0$$

The moment can be simplified largely. To eliminate the effect of the transformation, let

$$\tilde{f}(x, y) = f(x, y) - \frac{m_{0,0}}{k_a} - \frac{m_{1,0}}{k_b} x - \frac{m_{0,1}}{k_c} y$$

$$= k\{\hat{f}(x, y) - k_a^{-1}\hat{m}_{0,0} - k_b^{-1}\hat{m}_{1,0} - k_c^{-1}\hat{m}_{0,1}\}$$

Where $m_{0,0}, m_{0,1}, m_{1,0}$ are the moments of $f(x, y)$, and k_a, k_b, k_c are the constants designed to eliminate the first three order of the function $\tilde{f}(x, y)$.

Therefore $\tilde{f}(x, y)$ is k times of the transformed feature expression of $\hat{f}(x, y)$, which has been affected by an addictive plane $k_a^{-1}\hat{m}_{0,0} + k_b^{-1}\hat{m}_{1,0} + k_c^{-1}\hat{m}_{0,1}$ depending on the function itself. The parameter k can be any real number even is negative, which means the contrast is trivial to the shape analysis. Thus the invariance of the feature $\hat{f}(x, y)$ can be derived by normalize its moments vectors.

The invariant feature is derived from the invariant image $\tilde{f}(x, y)$, which is expressed with the original image moments.

$$u_{pq} = \int_{x_0}^{x_d} \int_{y_0}^{y_d} x^p y^q \tilde{f}(x, y) dx dy$$

$$
\begin{aligned}
&= m_{p,q} - m_{0,0}\frac{\int_{x_0}^{x_d}\int_{y_0}^{y_d} x^p y^q dxdy}{\int_{x_0}^{x_d}\int_{y_0}^{y_d} dxdy} \\
&\quad -m_{1,0}\frac{\int_{x_0}^{x_d}\int_{y_0}^{y_d} x^{p+1} y^q dxdy}{\int_{x_0}^{x_d}\int_{y_0}^{y_d} x^2 dxdy} - m_{0,1}\frac{\int_{x_0}^{x_d}\int_{y_0}^{y_d} x^p y^{q+1} dxdy}{\int_{x_0}^{x_d}\int_{y_0}^{y_d} y^2 dxdy}
\end{aligned}
$$

This moment is also a common Cartesian moment but performed on the retrieved invariant image, thus can be easily normalized and transformed into the other moments.

One advantage of the moments needed to be addressed is that the direction of the feature is also available at the same time according to its mass center offset.

$$
\begin{aligned}
x_m &= \frac{m_{1,0}}{m_{0,0}} \\
y_m &= \frac{m_{0,1}}{m_{0,0}} \\
\theta &= \angle(y_m - \frac{y_d + y_0}{2}, x_m - \frac{x_d + x_0}{2})
\end{aligned}
$$

Some experiments are carried out to search a finger model, the real image and similarity map are shown in the Fig.1. It can be seen clearly that the feature is very salient and the location is precise. Also it very efficient that over 10000 target are searched within a second. What's more, it can be largely improved by optimize the searching method as next section.

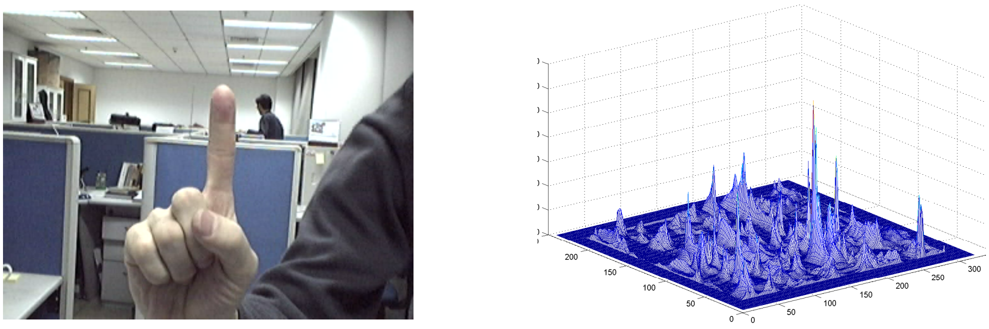

Fig. 1. The experiment of finger tip tracking. The left figure is the finger in the complex background; the right one is the matching result

3 Information aided searching strategy.

One of the basic problems in stereo vision is about detection and matching. In this section we design a simple algorithm to detect the moving salient features. Combining all the information and optimized by the filters, the complex matching procedure can be avoid.

Given a serial of images $I(x, y, t)$, to search a feature under some different scale s the matching procedure are expected to compute in a complexity of $O(x \times y \times t \times s)$. It's hard to be implemented in real time.

Fortunately, not all the location are needed to search and most of them are trivial. We propose a probability based searching strategy to found the most salient areas.

From the frequent space point of view, the frequency information is highly correlated with scales, and not all the frequency in the space is valuable, such as high frequency noise and low frequency lighting conditions.

$$P(I(x, y, t), x, y) = \int_s bf(s) \times ker(I(x, y, t), s) ds$$

Where the function $bf(s)$ is a band pass filter stress the certain frequency according to the feature scale searched. And $ker(\cdot)$ is a kernel based derivative filter, which present the certain frequency information. Here the canny methods are used to depress the near points so that some lines are selected to search under different scales(See Fig.2).

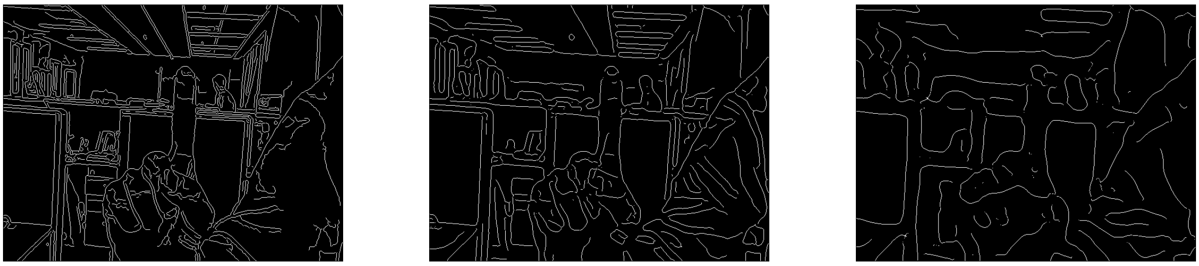

Fig. 2. The edge detected of different scales

Then from the time space point of view, in most cases the interactive subject are the few only moving object in the image, then the searching area can be reduced also by the probability of moving points. To avoid the instability of different image edges caused by noise, the moving probability is calculated independent with the spacial one. At last, the probability of all the space are considered together to form a searching path.

$$P(I(x,y,t),t) = \frac{dI(x,y,t)}{dt}$$

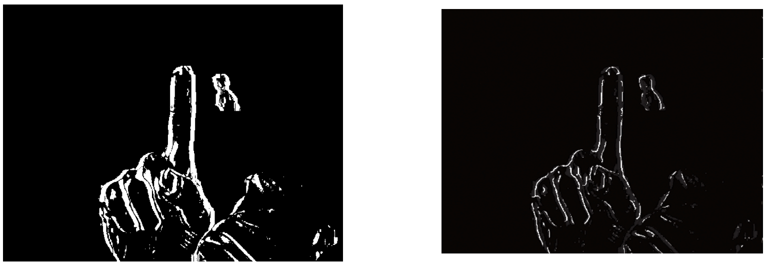

Fig. 3. Image difference and the integrate searching area

4 Multi Dimension of Freedom by Stereo Vision

When the features are detected from both of the cameras, the 3D data and the feature direction are to be retrieved by triangulation, before which the calibration should be done first.

Support a feature in the world coordinates x_w, y_w, z_w, are rigid transformed into the camera coordinates x_c, y_c, z_c by T_r and then perspective transformed into image space x_u, y_u by T_p

$$X_u = T_p X_c, X_c = T_r X_w$$
$$X_u = T_p T_r X_w = T_e X_w$$

Where, T_e is the extrinsic parameter of the camera that corresponding with the world coordinates, which need to be adjusted every time before the game start.

Another intrinsic parameter should not be omit, which will largely distort the image. Fortunately, for most cameras the intrinsic parameters are not changed and therefore can be just once and for all. Details of the method and parameters can refer to Zhang's[16].

Besides this, some new parameters are adopted to get more dimensional freedom with single feature. As mentioned above the direction of the feature is another important information for the object.

Support an object in the world space with the direction Q_w, then the projected direction of the two cameras are Q_{cl}, Q_{cr} respectively with the extrinsic transforms T_{el}, T_{er}

$$Q_{cl} = T_{el} Q_w \ , \ Q_{cr} = T_{er} Q_w$$

Given the extrinsic transforms and 2D vectors, the direction of the object can be evaluated by the intersect projected planes [17], which presented by the two angles, α angle with x axis and β angle with y axis.

However the result is not always accurate due to the noise and location result, an error analysis are made by the experiments. The average errors of the five parameters are listed below.

Table 1. Experiment error analysis of each parameters

Feature parameters	x	y	z	α	β
average error rate%	4.3	5.2	7.7	18.2	20.5

It can be seen from the table that not all the parameters are convincing, the x,y,z parameter are the most reliable variant that can be used as pointer in the 3D space, or a 2D plane with auxiliary control method such as speed range; The angle parameter are inaccurate due to the feature direction errors, however, it is suitable to give the binary direction control information in the game.

5 Experiment

A real-time system are implemented on the PC at 1.7GHz, it's efficient to track the finger in the three dimensional space with five DOFs (See Fig.4).

Also a game interface are tested with most accurate 3 DOFs, considering the requirement of the game control and the robustness. The architecture of the system is designed as Fig.5, and some sample of the game are showed in the Fig.6.

6 Conclusion

We proposed a new method to track features and up to 5 DOFs can be achieved from a single feature compared with the others, with is convenient to implement and to be used. The experiments show it is suitable for the interactive control of complex games. And more applications are prospected such as 3D mouse, 3D reconstruction and so on.

References

1. D. M. Gavrila: The visual Analysis of Human Movement: A Survey, Computer vision and Image Understanding. Vol 75, No. 1. (1999)
2. W.T. Freeman, K. Tanaka, J. Ohta, K. Kyuma: Computer vision for computer games.2nd International Conference on Automatic Face and Gesture Recognition, Killington, VT, USA. (1996)

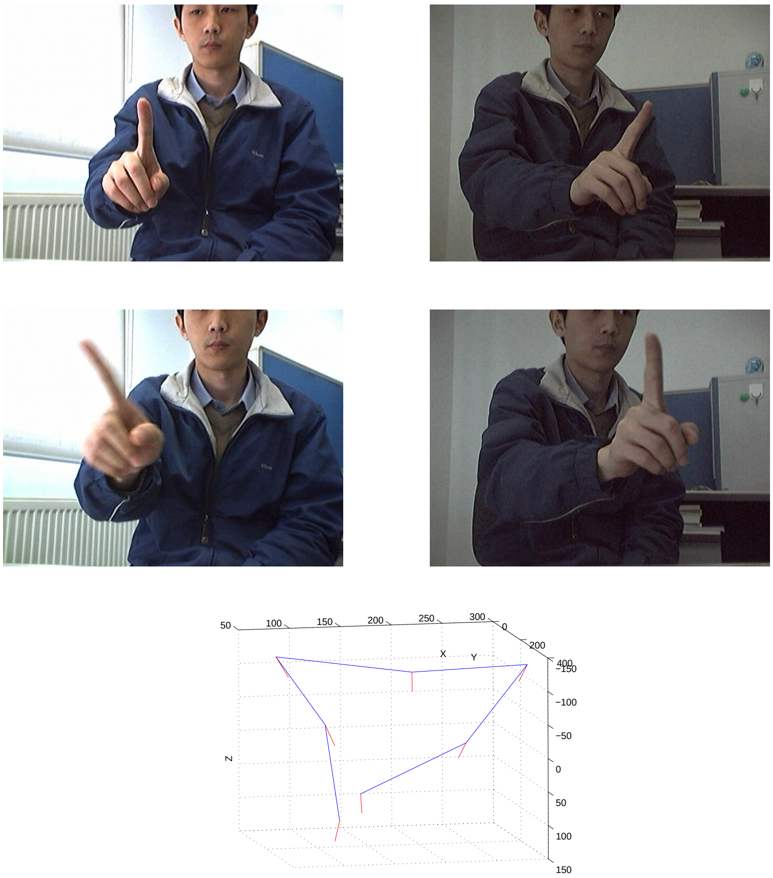

Fig. 4. Finger detection and its track in 3D

3. M. Betke,J. Gips, P.Fleming: The Camera Mouse: Visual Tracking of Body Features to Provide Computer Access for People With Severe Disabilities. IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 10, NO. 1, MAR (2002)
4. J. Magee, M. R. Scott, B. N. Waber, M. Betke: EyeKeys: A Real-time Vision Interface Based on Gaze Detection from a Low-grade Video Camera. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, (2004).
5. V. Pavlovic, Sharma and T. Huang: Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review, IEEE Transaction on Pattern Analysis and Machine Intelligence, VOL. 19, NO. 7, JULY (1997)
6. C. Fagiani. M. Betki. J. Gips. Evaluation of tracking methods for human computer interaction. IEEE Workshop on Application of Computer Vision. (2002)
7. Zhang, Z., Wu, Y., Shan, Y., and Shafer, S: Visual Panel: Virtual Mouse, Keyboard and 3D Controller with an Ordinary Piece of Paper. ACM Workshop on Perceptive User Interfaces, Nov. (2001).

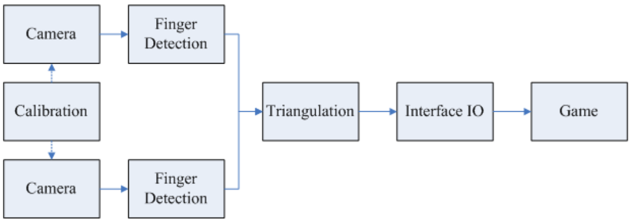

Fig. 5. Stereo vision game interface architecture

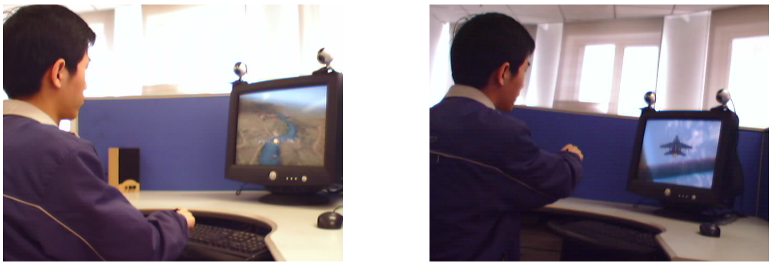

Fig. 6. Game played with stereo vision

8. D.M. Gavrila and L.S. Davis: 3d model-based tracking of humans in action: a multi-view approach, CVPR (1996)
9. R. C. Zeleznik, A. S. Forsberg and P. S. Strauss: Two Pointer Input For 3D Interaction, Proceedings of the symposium on Interactive 3D graphics, Providence, Rhode Island, United States (1997)
10. Y. Wu, K. Toyama and T. S. Huang: Self-Supervised Learning for Object Recognition Based on Kernel Discriminant-EM Algorithm, in Proc. IEEE Int'l Conf. on Computer Vision (ICCV'01), Vol.I, 275-280, Vancouver, Canada, July, (2001)
11. Sidenbladh, M. Black and D. Fleet: Stochastic Tracking of 3D Human Figures using 2D Image Motion, ECCV (2000)
12. M-K. Hu: Visual pattern recognition by moment invariants, IRE Trans. on Information Theory, IT-8:pp. 179-187, (1962)
13. J. Heikkila: Pattern matching with affine moment descriptors, Pattern Recognition, 37, pp: 1825C1834, (2004)
14. C. Chong, P. Raveendranb, R. Mukundanc: Translation invariants of Zernike moments, Pattern Recognition, 36, pp:1765C1773,(2003)
15. R. Takamatsu, M. Sato, H. Kawarada: Pointing device gazing at hand based on local moments, Proceedings of SPIE , Volume 3028 Real-Time Imaging II, pp. 155-163,April (1997)
16. Z. Zhang:Flexible Camera Calibration By Viewing a Plane From Unknown Orientations, ICCV (1999)

17. R. I. Hartley, P. Sturm: Triangulation,Computer Vision and Image Understanding, page 957-966, (1994)