DEPARTMENT OF
APPLIED PHYSICS AND ELECTRONICS
UMEÅ UNIVERISTY, SWEDEN

DIGITAL MEDIA LAB

Is A Magnetic Sensor Capable of Evaluating A Vision-Based Face Tracking System?

Zhengrong Yao [1]
Dept. Applied Physics and Electronics
Umeå University
SE-90187, Umeå Sweden
e-mail: Zhengrong.Yao@tfe.umu.se

Haibo Li [2]
Dept. Applied Physics and Electronics
Umeå University
SE-90187, Umeå Sweden
e-mail: haibo.li@tfe.umu.se

DML Technical Report: DML-TR-2003:6
ISSN Number: 1652-8441

Report Date: December 20, 2003

1
2

Abstract

This paper addresses an important issue, how to evaluate a vision-based face tracking system? Although nowadays it is getting popular to employ a magnetic sensor to evaluate the performance of such systems. The related issues such as condition and limitation of usage are often omitted. In this paper we studied this accepted evaluation methodology together with another evaluation method, Peak Signal to Noise (PSNR) commonly used in image coding community. The condition of proper usage of magnetic sensor as evaluating system is discussed. Our theoretical analysis and experiments with real video sequences show that we have to be very careful to select the so-called ”ground truth”. We believe that to help further development of face tracking techniques, a valid performance evaluation is necessary, both the evaluating system and the tracking system have to be jointly considered to decide if the evaluating method is valid. The experimental results give us further hints about the tracking performance when using different tracking scheme.

Keywords

Universal Multimedia Access, Content adaption, Model based coding (MBC), Analysis by synthesis, Simulated Annealing.

1 Introduction

Vision based face tracking has its potential usage in many applications such as Virtual Reality, video surveillance, teleconferencing, Human computer interaction, etc. The key challenge is to extract the 3D motion information of the face object in video. Face tracking with a monocular camera has been recognized as a hard computer vision problem, which has attracted a lot of researchers from different research communities such as computer vision, image coding and computer graphics, to work on it. Although there is no widely accepted solution for this problem yet, the 3D wireframe model-based face tracking technique has been accepted as the most promising solution [4, 5, 9]. The key idea is that a generic 3D geometric face model is used to "regularize" facial motion estimation [10, 5, 6]. The tracking methods which implement a geometrical model are often referred as model-based face tracking.

One important issue in designing a vision based face tracking system is how to evaluate its performance. A reasonable evaluation scheme could both reveal the tracking performance and also give hints on how to improve the design. Depending on what is the concern, different application has different evaluation method. Two representative evaluation methods for motion tracking have been widely used. One is the method used in video coding community, which use the "Peak signal to noise ratio" (PSNR) curve as the common evaluation method, the other one is used in computer vision community, which use the "ground truth" measurement as reference for evaluation of tracking performance. The comparison of two methods is list in Table I

Method	PSNR	Ground truth
Requirement	Original image and reconstructed image	Ground truth got from another measurement source
Work space	image	motion parameters
Accuracy	Pixel level	Varying according to application
Evaluation basis	Visual effect	parameters bias

Table I: Comparison of evaluation with PSNR method and ground truth method

It could be seen that PSNR compares two images where ground truth method compares two set of parameters. PSNR method revel the effect of pixel changes, which in turn caused by parameter changes. Thus PSNR method is related to ground truth method, but it does not necessarily reveal the true changing of parameters. PSNR curve thus can not give direct performance evaluation for the motion tracking. The ground truth method give direct evaluation on the performance of motion tracking, it is heavily depended on an available high accuracy measurement system which provide the "ground truth". In computer vision community, a widely used way is to compare the visual tracking result with a highly accurate motion sensor such as a magnetic motion tracker [5, 6] which suppose to be able to be the ground truth provider. Depending on different application, the accuracy of the required ground truth data is different. i.e. Virtual Reality (VR) often need very high accuracy to prevent spatial confliction of virtual figures. Model-based coding (MBC) often need the tracking accuracy at pixel level, where not care much about real spatial accuracy. Human computer interaction (HMI) often need lower accuracy where approximate motion information is good enough. In this paper, we choose to use MBC application as an example due its potential usage and its moderate requirement of accuracy. We show in this paper that for a MBC application, the commonly accepted evaluation method of using a magnetic sensor often does not fit for the evaluation task. For convenience of discussion, we briefly review the concept of MBC.

1.1 Model-based coding (MBC)

MBC is a promising video coding technique targeted for very low bit rate video transmission [1, 2, 3, 4]. The idea behind this technique is to parameterize a talking face based on a 3D face model. Parameters describing facial movements and texture information are extracted at the transmitter side. The extracted parameters are then sent to the receiver to synthesize the talking head. Very high compression is achieved since only high-level, semantic motion parameters are transmitted. Figure 1 shows a basic block diagram of a MBC system. The incoming video frame is subjected to analysis process in encoder, where the scene parameters are extracted. The decoder helps the synthesis process for receiver.

In this paper the focus is put on the performance evaluation for the 3D facial tracking issue, which happens in the encoder side. The facial tracking in a MBC system is implemented in two stages: initialization and successive

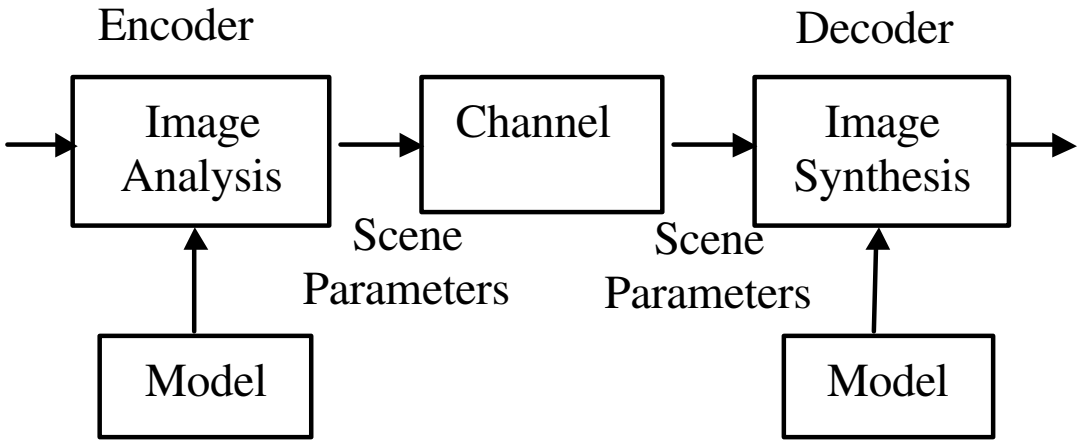

Figure 1: The block diagram for a Model-based coding system.

tracking. Initialization is the process of fitting a generic 3D face model to a target face in the first video frame.The facial texture is also extracted during initialization. In the successive tracking stage, the face model is made to follow the moving head by automatically recovering the face position and the facial expression from each frame in the video sequence. The task of *model based face tracking* is : to fit a generic 3D wireframe model to a target face. The fitting should try to keep the largest consistency of the mapping from model point **P** to face image point **P'**, this is shown in Figure 2. The task of *performance evaluation* for the model-based face tracking is: to evaluate the accuracy of the tracking result, which at final reflected in the quality of the synthesized video at the receiver side. PSNR is often served as the evaluation tool through comparing the original video frame with the synthesized one. Note that PSNR curve is also affected by many other factors other than pixel changes such as the texture extraction accuracy, illumination changing, synthesis errors, etc.

2 Magnetic tracker and visual tracker

In order to evaluate the performance of a motion tracking system, suppose we always need to rely on using another evaluating system. The evaluating system could refers to any kind of location sensing system [7]. The requirement of the evaluating system for the specific tracking system is dependent on the application field. A basic requirement is: the evaluating system should be more accurate than the tracking system, to give a sound evaluation for the tracking performance.

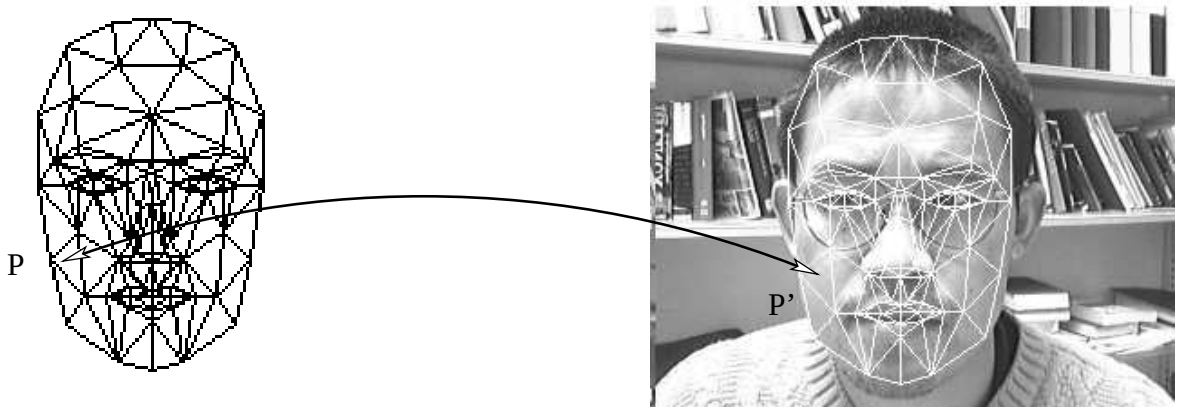

Figure 2: The model based face tracking: fit a generic 3D model onto one face in video frame.

Figure 3 shows a general configuration for the task of performance evaluation for a visual tracking system. To facilitate the discussion, the components in the system are labelled with different coordinate system. There

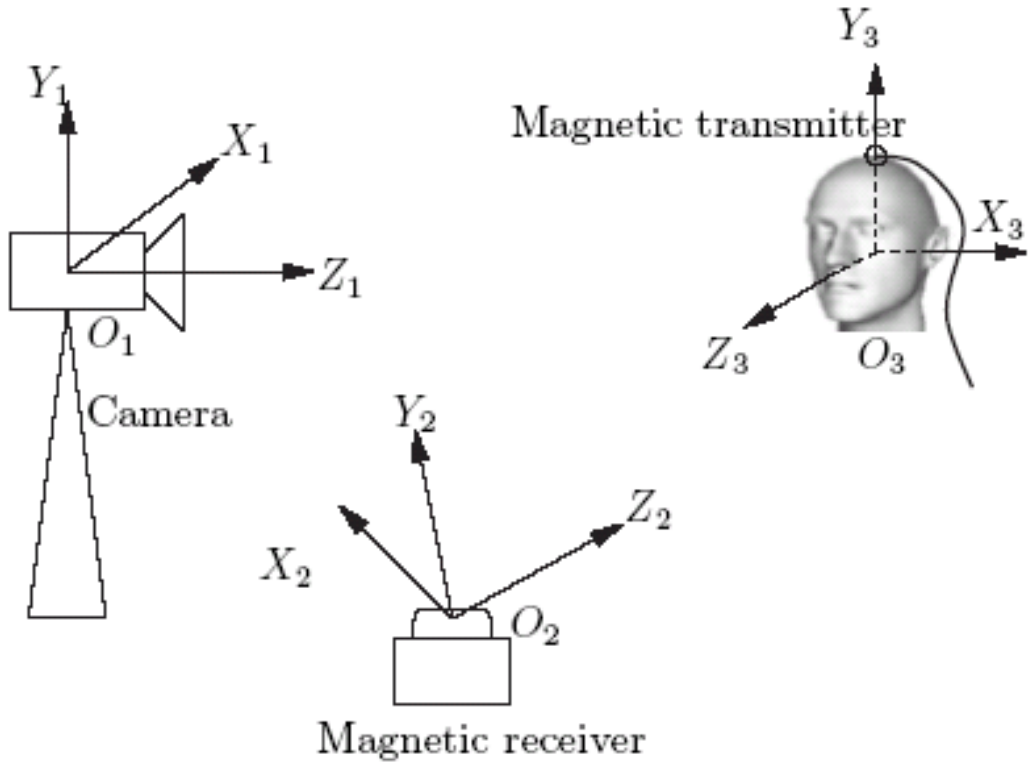

Figure 3: Three coordinate system in performance evaluation.

are commonly three coordinate system. One is the camera coordinate system which is used in the visual tracking. The second one is the evaluating system, which in our discussion is the magnetic tracker. A third one is the local object coordinate system, which in our case is the targeted face object coordinate system. The origin and three axes of the coordinate systems are labelled as O_i, X_i, Y_i, Z_i, where $i = (1,2,3)$ refers to camera, evaluating system, local object system respectively. In our discussion, the magnetic tracker (the evaluating system)is supposed to be the responsible one for providing the ground truth data for evaluation task. The magnetic tracker measures the magnetic transmitter's position and orientation. The magnetic transmitter is often attached to the face object. In Figure 3 the magnetic transmitter is put on top of head and indicated by a circle.

A question is then: what is a good evaluating system for a MBC system? As discussed before, the requirement for the evaluating system is its measure accuracy. Here we give the empirical estimation of the required accuracy for the evaluating system for a MBC system.

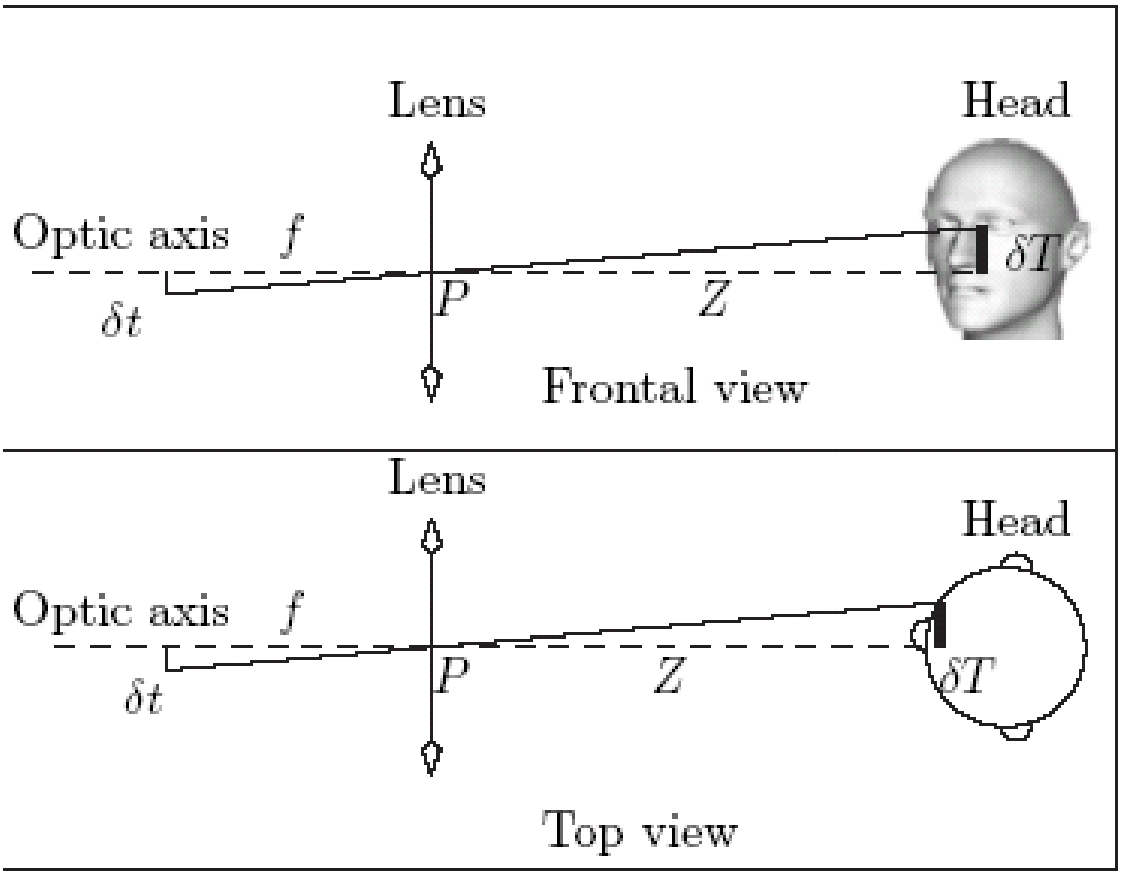

Figure 4: Accuracy requirement estimation for the performance evaluation in MBC. Upper row: estimate position accuracy; Lower row: estimate orientation accuracy.

Figure 4 shows the pinhole projection model of the camera system. We can estimate the measurement accuracy requirement for a face tracking system as shown in Figure 4. The distance t on image plane of the distance

on object T are related by Equation 1:

$$t = f\frac{T}{Z} \tag{1}$$

where Z is depth and f is focal length. We thus have:

$$T = \frac{tZ}{f} \tag{2}$$

let δt denote the measurement error of T,

$$\delta T = \frac{\delta t Z + t\delta Z}{f} \tag{3}$$

we can then have the relative measurement error of value T

$$\frac{\delta T}{T} = \frac{\delta t Z + t\delta Z}{fT} = \frac{\delta t}{t} + \frac{\delta Z}{Z} \tag{4}$$

it is related to the measurement error δt in image plane. If the image plane has the sub-pixel accuracy, that is $\delta t = \frac{L}{2}$ where L denote the image resolution ($L = 1 pixel$). It is common that $\delta Z \ll Z$, thus $\frac{\delta Z}{Z} \approx 0$. The relative measurement error could be written as:

$$\frac{\delta T}{T} = \frac{Z\delta t}{fT} = \frac{ZL}{2fT} \tag{5}$$

as a real example, for a Philip web camera with f 3mm lens, the distance $Z = 30$ inches, f is approximate 330 pixel for image resolution of 320×240. The head size $T = 10$ inches, relative measurement error is $\frac{30\times 1}{2\times 300\times 10} = 0.0017$, this means that if the head size is 10 inches, we can expect the measurement error as $10 \times 0.0017 = 0.017$ inch. The evaluating system accuracy should has at least ten time less of the measurement error, that is 0.0017 inch .

For orientation measurement (see top view of Figure 4), since the evaluating system only need to measure out the angle which affect the pixel changes. Thus the translation measure accuracy decide the angle measurement accuracy. Suppose the radius of head is around 5 inches, the angle accuracy is thus $\frac{0.0017}{5} = 0.00034$ radian, or 0.019 degree. It could be seen that for such a common environment and devices setting, we need rather high requirement for the evaluating system.

Suppose such a evaluating system is available, we then need to compare the measurement value with the vision based tracking results. As shown in the Figure 3, the visual tracker and the magnetic tracker work in two different coordinate system. A calibration procedure is inevitable before comparing the two sets of tracking result from these two systems. To evaluate the tracking outcomes from a model based face tracker, the two set of tracking result need to be converted into one same coordinate system. In [5] this is treated as an *absolute orientation problem*, which aims at finding the transform matrix given two cluster of 3D points from two different coordinate systems.

3 Calibration using solution for an absolute orientation problem

Suppose the 3D points of face object in the magnetic system are denoted as $\mathbf{S} = (x_i, y_i, z_i)^t, i = 1, ..., n$. The corresponding camera-space points are denoted as $\mathbf{S}' = (x_i', y_i', z_i')^t$. The two coordinates are related by a rigid transformation which contains rotation $\mathbf{R}$ and translation $\mathbf{T}$:

$$\mathbf{S}' = \mathbf{R}\mathbf{S} + \mathbf{T} \tag{6}$$

where

$$\mathbf{R} = \begin{pmatrix} \mathbf{r}_1^t \\ \mathbf{r}_2^t \\ \mathbf{r}_3^t \end{pmatrix} \tag{7}$$

and

$$\mathbf{T} = \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \tag{8}$$

is the rotation matrix, **T** is the translation matrix, When the non-rigid motion is small and negligible, the transform in Equation 6 thus has only rigid translation and rotation. This is often true for using calibration sequence which contains only rigid pure rotation or pure translation. In this case the problem changes into the absolute orientation problem. The mapping from the points **S** onto image plane point $\mathbf{V} = (u_i, v_i, 1)^t$ is through a projection:

$$u_i = f\frac{\mathbf{r}_1^t \mathbf{p}_i + t_x}{\mathbf{r}_3^t \mathbf{p}_i + t_z} \tag{9}$$

and

$$v_i = f\frac{\mathbf{r}_2^t \mathbf{p}_i + t_x}{\mathbf{r}_3^t \mathbf{p}_i + t_z} \tag{10}$$

where f refers to the focal length and is supposed to be known. Different method have been proposed for solving the absolute orientation problem. [8] gives both a good review of related works and also propose a fast solution. Through calibration, a good transform could be found for a reference frame. The motion estimation could be treated as the relative motion estimation between the reference frame and current frame. Since the calibration for the reference frame is available, the comparison of the visual tracking results and ground truth is possible. [6] follows similar way. Although this evaluating scheme has sound theoretic background, [11][12] shows that the magnetic tracker often need intensive calibration work to make the measurement meaningful. We will also show its deficiency when applied this to the MBC system.

4 Deficiency of using magnetic tracker for evaluating visual tracker in MBC

When applied in MBC system, a detailed study shows that there is deficiency in using the magnetic sensor as the evaluation system. To show this fact, we used the real image sequence and the ground truth data used in [5, 6] which is available at http://www.cs.bu.edu/groups/ivc/HeadTracking. The ground truth head motion were got by "Flock of Birds" 3D tracker. As also noted in [5], the magnetic tracker is subject to measure distance, metal presence and other factors. Although the nominal positional accuracy is 0.1 inches and angular accuracy of 0.5 degree. In practical the magnetic measurement shows large jitters. the noise level is much larger than the nominal accuracy. Figure 5 shows the histogram of the typical jitters happened in a translation x_i sequence. The jitters happened in other parameter measurements are similar with what shows in Figure 5. From the figure it is clear that jitters at range of 0.2 to 0.3 inches are quite common. Through calibration, an estimation shows that these jitters correspond to a jitter in range of 2 to 3 pixels respectively. Although the range could change due to depth changing and also face object size, these range of jitters are obviously far too large to serve as "ground truth" in a MBC system. This is because when talking about the video coding performance, we always work on the pixel accuracy level. A 2 to 3 pixels ambiguity will not lead to large degradation of visual quality but it will lead big change in PSNR.

We conduct a experiment to show the effect of treat the magnetic tracker as "ground truth". Since the jitters exist in the measurement data, the only way we could made use of the data is to preprocess it, trying to get rid of the effect of jitters. After carefully "smooth" and data, we could use the solution for absolute orientation problem or fine manually adjustment to find the transform between two coordinate system for the reference frame. The face model is then driven by the "ground truth" and texture-mapped to synthesize the video. A PSNR curve is also estimated through out the video sequence as shown in Figure 8.

In this case, we have only the magnetic data. Thus there is no real "ground truth". The only possible comparison we could perform is with manually fitting data according to subjective judgement of the matching effect.

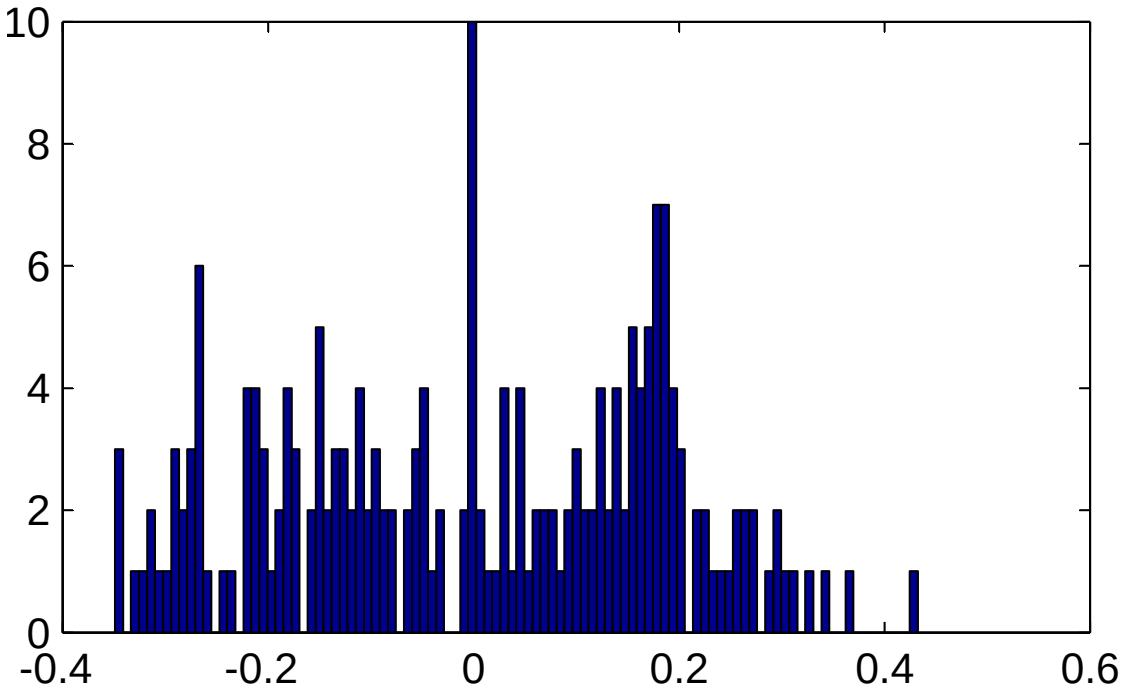

Figure 5: Typical measurement jitters for translation x_i.

We carefully manually fit the face model onto each frame in the testing video sequence. Then the face model was driven by both the ground truth data and manually fitted data and the PSNR curve is estimated (Figure 8).

We also use a simple vision based tracking scheme to tracking the face. We use the texture extracted at the first frame, and use Simulated Annealing (SA) scheme to perturb the motion parameter and use these motion parameters to drive the face model to check the matching between the original image and the texture-mapped one. The searching stop when certain number of loop is reached.

Figure 6 shows the synthesized video from the ground truth data and the vision based tracking result. Figure 8 shows all the PSNR curves of reconstructed from manual fitting data, magnetic tracking data, vision based tracking data.Figure 7 shows the parameters estimated from the vision based tracking system and the magnetic sensor. For the vision based tracking, two set of curves are shown in the Figure 7. One is the vision based tracking in each frame. One is the smoothed version of the individual tracking result just for illustrating the trend of the curve.

Note the PSNR is calculated over the whole video frame with the synthesized face, although in practice the background images are not available. Since the synthesized face is overlapped on the original video frame to calculate the PSNR curve, the PSNR trend is not affected when background image is not available, thus still give us the right picture in term of comparison of the two cases.

4.1 Experimental result and discussion

From Figure 6 and Figure 8, it could be seen that the vision based tracking and manual fitting by subject evaluation often give a better ”explanation” of the face image compared to the magnetic tracking. The PSNR produce with magnetic tracking data is worse compare to the manually fitted data and vision based tracking data. An estimation

shows that for this sequence, a gap of almost 5 dB degradation is found.

A vision based system could get even better PSNR curve, yet it does not necessarily give more accurate visual tracking result. What's more, through comparing in Figure 7, it says that the vision based tracking system could have different explanation as the magnetic sensor. From Figure 7 it could be seen that manual tracking show different trend as magnetic tracking at some part of the testing video, the vision based tracking basically follows the trend of manual fitting. Together with the subjective evaluating of the reconstructed video using three kinds of tracking results. We come to the conclusion that manual fitting give the best confidence in terms of accuracy of motion parameters extraction, while vision based tracking give best PSNR curve, the magnetic tracking give worst results in both cases.

From the theoretical analysis we have reached the conclusion that, since the accuracy of the magnetic sensor does not match our requirement for vision based tracking. Our experiments rectify our conclusion, what's more, we gain further hints about the performance of vision based tracking when using different tracking scheme.

Our vision based tracking system is appearance based method. In comparison, [10] use feature point based tracking system and can have even higher accuracy than magnetic sensor, in that case, the use of magnetic sensor as evaluating system make little sense. In fact, our manual fitting is also based on facial feature fitting, the experimental result convince us that the manual fitting is most trustworthy. For appearance based method, the magnetic sensor does not always provide trustworthy ground truth data. This also show that for explaining image, the appearance based method could work well, but for estimating 3D motion parameters, the appearance based method could perform worse compared with feature based methods.Depending on what kind of vision based face tracking system, the use of magnetic sensor as evaluating system is sometimes valid, sometimes not.

evaluation using magnetic sensor?	In theory	In practical
Feature based visual tracking	Yes	No
Appearance based visual tracking	Yes	No(Yes)

Table II: The conditions of using of magnetic sensor as evaluating tools.

Table II summarize these knowledge got from our experiments. It says that when the vision based tracking system use feature based method, both in theory and in practical the magnetic sensor could not match the accuracy requirement. If the tracking system use appearance based method, since they often have less tracking accuracy compare to feature based method, thus in theory the magnetic sensor could serve for evaluating purpose, but in practical it has to be evaluate first if the accuracy of magnetic sensor is worse then the appearance based tracking system. The conclusion could be yes or no.

From the individual frame visual tracking we can see that the "jitter" is even bigger than the magnetic sensor, this tells us that without a motion model constraint, the appearance based visual tracking only try to find the "best" explanation for the image, it cannot guarantee a smooth visual effect for the tracking result. Thus for appearance based method to work fine, a motion model constraint is still needed.

5 Conclusions

We have to be very careful to select the so-called "ground truth", when evaluating the performance of a vision based face tracking system. We believe that to help further development of face tracking techniques, a valid performance evaluation is necessary, both the evaluating system and the tracking system have to be jointly considered

to decide if the evaluating method is valid. In this work, we discuss only the MBC, whose central task for tracking is to give a good explanation for video images, we believe this is representative in vision based applications and the conclusion for using magnetic sensor could be also valid for other vision based applications.

Acknowledgments

We wish to thank Marco La Cascia and Jing Xiao for their kind reply of our questions, Robert Forchheimer and Lena Klasén for the discussion.

References

[1] R. Forchheimer, O. Fahlander, and T. Kronander, ”Low bit-rate coding through animation,” *Proc. International Picture Coding Symposium PCS‘83*, pp. 113–114, Mar. 1983.

[2] K. Aizawa,H. Harashima, and T. Saito, ” Model-based analysis synthesis coding system for a person’s face,” *Signal Processing: Image Communication*, Vol. 1,No. 2, Oct. 1989.

[3] W.J. Welsh, ”Model-based coding of moving images at very low bit rates,” *Proc. Int. Picture Coding Symp.*, Stockholm, Sweden, 1987.

[4] H. Li, ”Low Bitrate Image Sequence Coding,” Ph.D Thesis, Linkoping, 1993.

[5] M. L. Cascia, S. Sclaroff and V. Athitsos, ”Fast, Reliable Head Tracking under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D Models,” *IEEE PAMI*, Vol. 21, No.6, June 1999.

[6] J. Xiao, T. Moriyama, T. Kanade, and J. Cohn, ”Robust Full-Motion Recovery of Head by Dynamic Templates and Re-registration Techniques,” *International Journal of Imaging Systems and Technology*, September, 2003.

[7] Jeffrey Hightower and Gaetano Borriello,”Location Systems for Ubiquitous Computing,”*IEEE Computer*,Vol.34, No.8,August, 2001.

[8] Chien-Ping Lu, Gregory D. Hager and Eric Mjolsness, ”Fast and Globally Convergent Pose Estimation from Video Images,”, *IEEE PAMI*,Vol.22,No.6,2000.

[9] S. Basu, I. Essa, and A. Pentland, ”Motion regularization for model-based head tracking,” *ICPR ’96*, Vienna, Austria, 1996.

[10] A. Azarbayejani and T. Starner and B. Horowitz and A. Pentland, ”Visually Controlled Graphics,”*IEEE PAMI*, Vol.15,No.6,1993.

[11] V. Kindratenko, and A. Bennett, ”Evaluation of Rotation Correction Techniques for Electromagnetic Position Tracking Systems,” in *Proc. Virtual Environments 2000 Eurographics Workshop*.

[12] G. Zachmann, ”Distortion correction of magnetic fields for position tracking,” *Proc. Computer Graphics International*, pp. 213–220, Belgium, June 1997.

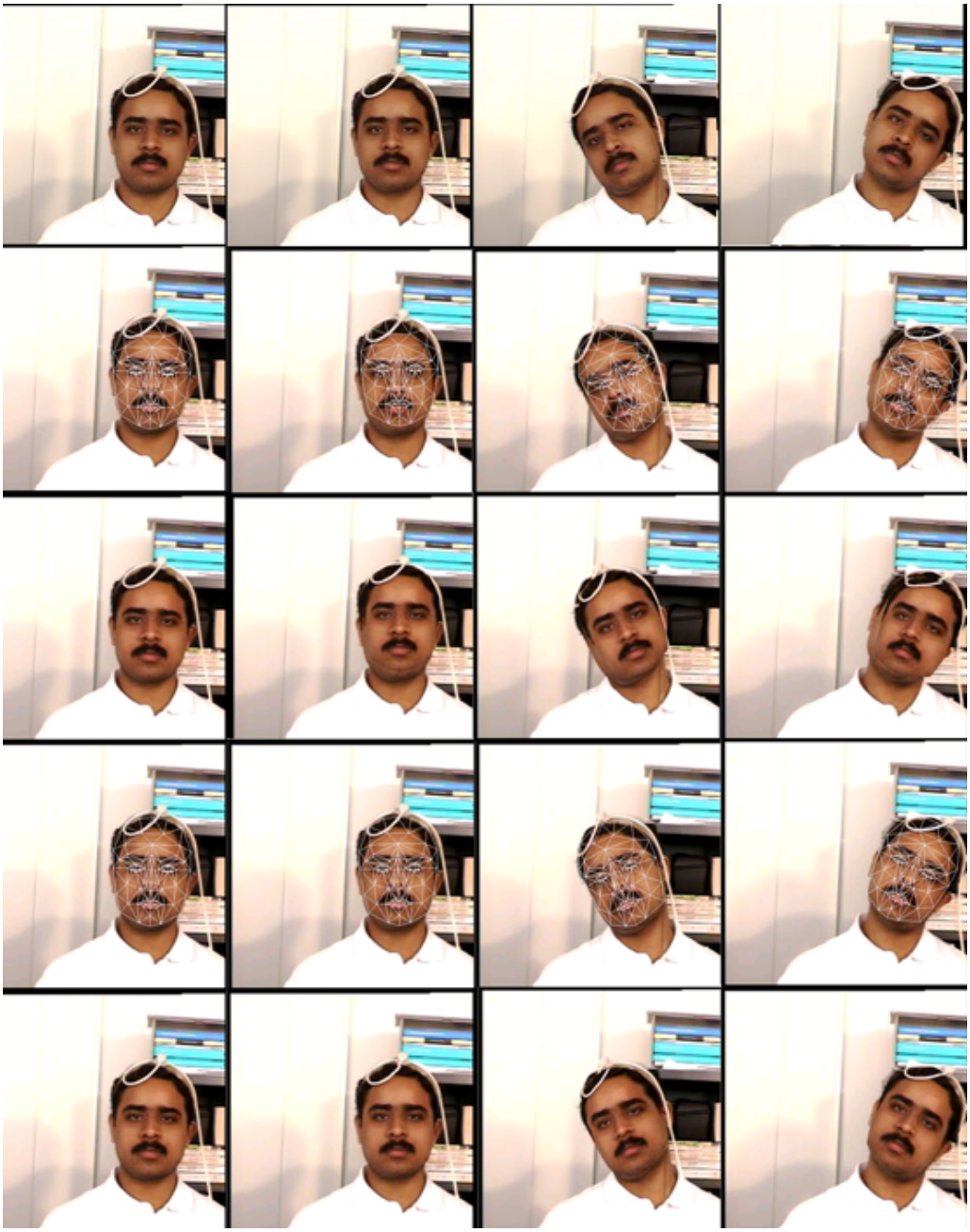

Figure 6: Row 1: original video frame 1, frame 20, frame 60, frame 140; row 2: synthesized face model from ground truth data overlapped on upper frame; row 3: synthesized video with texture face model. row 4: estimated face model from vision based tracking system, row 5: synthesized video with texture face model..

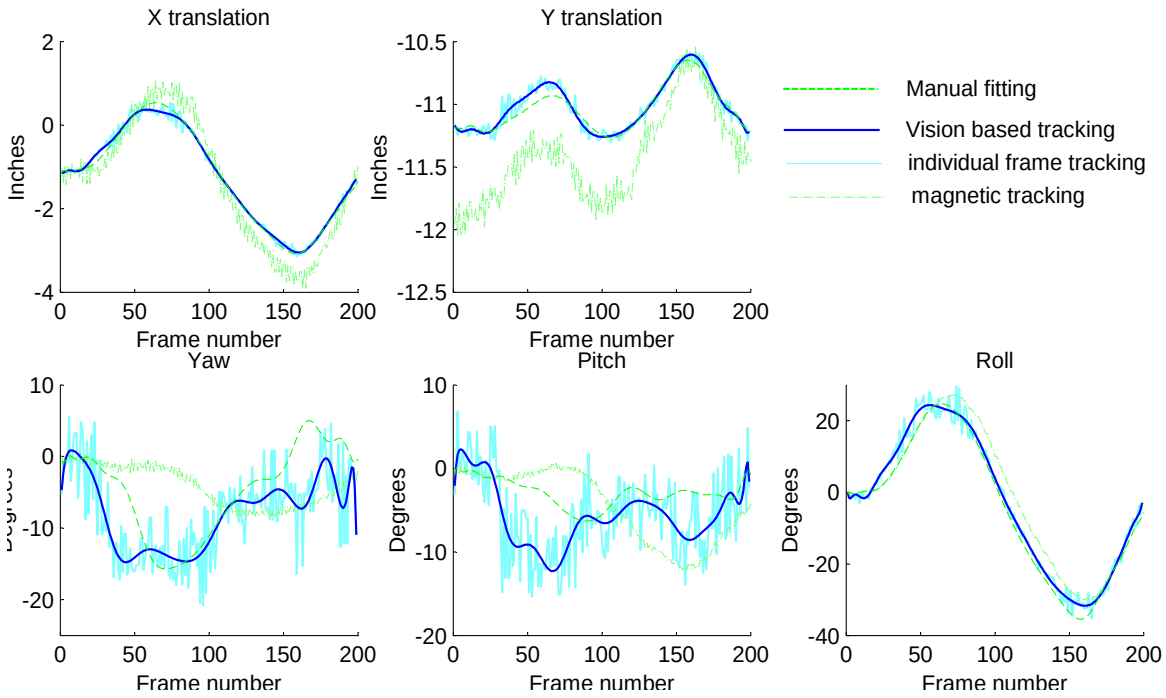

Figure 7: Motion parameters estimated from the global vision based face tracking system (thick solid line), individual frame tracking (thin solid line), magnetic sensor (dash doted line), manual fitting (dashed line). The translation in Z direction is not include since in vision based tracking system the parameter of scaling factor is used instead of depth Z.

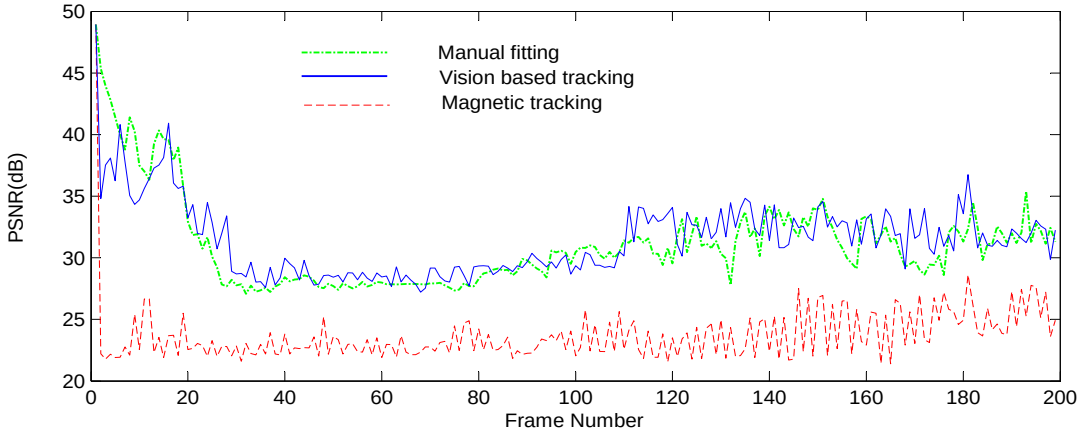

Figure 8: PSNR curves for video reconstructed from the vision based face tracking system (solid line) and got from the magnetic sensor (dash doted line) and manual fitting (dashed line).