

On Using Gait Biometrics to Enhance Face Pose Estimation

Sung-Uk Jung and Mark S. Nixon

Abstract—Many face biometrics systems use controlled environments where subjects are viewed directly facing the camera. This is less likely to occur in surveillance environments, so a process is required to handle the pose variation of the human head, change in illumination, and low frame rate of input image sequences. This has been achieved using scale invariant features and 3D models to determine the pose of the human subject. Then, a gait trajectory model is generated to obtain the correct the face region whilst handling the looming effect. In this way, we describe a new approach aimed to estimate accurate face pose. The contributions of this research include the construction of a 3D model for pose estimation from planar imagery and the first use of gait information to enhance the face pose estimation process.

I. INTRODUCTION

Calculating the 3D head pose is a fundamental process for unconstrained automatic face recognition systems. The 3D head pose describes not only direction but also basic information such as the size and position of the head. Accurate information of the head pose is also essential for face recognition.

Basically, estimating the head pose requires calculation of the translation and rotation information in 3D space. There are two main categories in existing methods: using an actual 3D head model or using an approximated head model. In the case of using actual 3D head models such as 3D Active Appearance Model (AAM) [1] and 3D Morphable Model (3DMM) [2], the advantage of these models is to be able to obtain accurate 3D texture and face shape. However, there are some disadvantages such as exact initialization is required and imperfect initialization can cause tracking errors. Unlike the actual models, the approximated head models such as 2D plane, 3D cylinder, and 3D ellipsoid cannot generate the actual shape and texture of a face. However, the model can be simple to implement and the computational load of a fitting process is much smaller than the actual model methods. Also, the initialization can be automatic.

To estimate 3D head pose using an approximate head model Liu et al. [3] used the Scale Invariant Feature Transform (SIFT) [4] to match the corresponding feature points between two adjacent views. Using Epipolar geometry [5], the fundamental matrix was calculated to convert the fundamental matrix into the essential matrix to obtain the pose information. Hager et al. [6] generated a 2D plane model using a single camera and Lucas-Kanade tracking. Cascia et al. [7] generated a 3D cylinder model. 3D head motion was treated as a linear combination of motion templates and

orthogonal illumination templates. Basu et al. [8] interpreted the optical flow in terms of the possible rigid motion and applied it to heads with a variety of shapes and hair styles, using a 3D ellipsoidal model. Xiao et al. [9] used the 3D cylinder model to track the head. To fit the face into the model Iteratively Re-weighted Least Squares (IRLS) technique was adapted. Gaurav et al. [10] showed 3D pose estimation using a particle filter without known camera focal length. Jang and Kanade [11] used SIFT and normalized correlation methods to extract and match the feature points.

There are three main approximate head models. The 2D plane model is simple but not effective for the human head since it does not represent curved surfaces and is not robust to out-of-plane rotations. The 3D cylinder model can represent a vertically curved surface well. However, it is a less accurate model than ellipsoidal model because the ellipsoidal model considers the horizontally curved surface as well. So, we shall use the 3D ellipsoidal model to represent the human head.

In this paper, we focus on the specific but one of the most general environments in visual surveillance where the subject is viewed walking towards a camera. We show that gait information can be used to exploit the looming effect for face localization. We suppose that the image sequences are recorded at low frame rate and the resolution of the face image varies according to the distance between the person and camera. To adapt to a low frame rate, a 3D ellipsoidal head pose model is built, and using the non linear optimization method the system can estimate 3D head pose. However, the head pose model can fail to track the head when there is large head pose variation and change in illumination.

Thus, we use additional information – the gait trajectory. There have been studies on fusing face with gait for recognition [12-14], but these have not used gait to mediate face acquisition.

To use gait characteristics, each frame gait trajectory can be obtained from the movement of the corresponding points between frames. Analyzing the gait trajectory enables the system to detect approximate face regions. In these face regions the detection rate can be increased. Figure 1 shows a data capture environment – the biometric tunnel [15][16] which uses eight synchronised IEEE1394 cameras to be able to capture gait, face, and ear images. The resolution of a face image is 1600×1200 pixels image [17]. The database used here is images taken from a frontal face camera as a person walked through the biometric tunnel and we can see that the resolution of the face is low at the start and moderate at the end of tunnel.

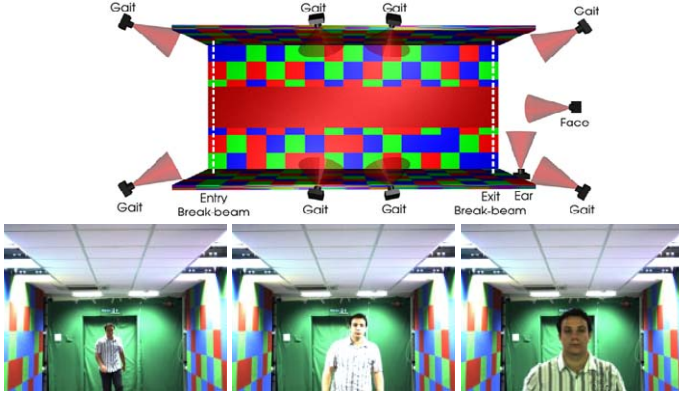


Fig. 1. Biometric tunnel and the sample front view data

II. 3D HEAD POSE ESTIMATION

If the 3D corresponding points exist and there is a small movement between images, the motion vector can be calculated by using the twist representation [18] or optical flow. Another way is to use epipolar geometry [5] to obtain the essential matrix from which the rotation and translation information can be extracted. However, the first calculates the motion vector based on the angle approximation in 3D space. The second can work in the case that the texture of image are clear so the optical flow can be calculated (it could not work properly especially in the low resolution image). The third method could be unstable if the corresponding points used in calculating fundamental matrix are on the same plane. All of the above methods are not suitable for our application. Therefore, we use another way to estimate the head pose. First of all, we shall define the objective function which describes relationship between reconstructed 3D points and 2D corresponding points. Then, we shall calculate the motion vector by using non-linear optimization.

A. 3D Point Reconstruction using a 3D Ellipsoidal Model

We use a 3D partial ellipsoidal representation [19]. Unlike the cylinder model the partial ellipsoidal model only considers the distinguishable face part, excluding the top and the bottom of a head, ears and background. The 3D partial ellipsoidal model is defined by following:

The one point on the 3D object (\mathbf{P}_0) can be represented by $[X_0 \ Y_0 \ Z_0]$. where $X_0 = r_x \sin \alpha \sin \beta$, $Y_0 = r_y \cos \alpha$, $Z_0 = r_z \sin \alpha \cos \beta$, r is the radius along each axis, α and β are axis angles. To find corresponding points between two images, Speed-Up Robust Feature (SURF) [20] is deployed.



(a) Model fitting Sample (b) SURF in 1st frame (c) SURF in 2nd frame
Fig. 2. Ellipsoidal model fitting image and matched SURF points

We aim to calculate a 3D rotation and translation matrix from the SURF points matched between the adjacent images. However, basically, it is impossible to extract the 3D information from 2D corresponding points. So, we use an alternative way of reconstructing position of points in 3D space using a 3D ellipsoidal modal.

Assuming the initialization has done in the first frame 3D points of SURF can be reconstructed by using the following method. Let point $\mathbf{p}_{3d} = [p_{3d_x} \ p_{3d_y} \ p_{3d_z}]$ in 3D space

$$\begin{aligned} \mathbf{p}_{3d} &= [\mathbf{R}^{-1}(\mathbf{K}^{-1} \mathbf{p}_{2d} - \mathbf{T})] \\ \alpha &= \cos^{-1}(p_{3d_y} / r_y) \\ \beta &= \sin^{-1}(p_{3d_x} / r_x / \sin(\alpha)) \\ p_{3d_z} &= r_z \sin(\alpha) \cos(\beta) \end{aligned} \quad (1)$$

where $\mathbf{R}, \mathbf{T}, \mathbf{K}$ represent the rotation, translation, camera matrix, \mathbf{p}_{3d} is a reconstructed 3D SURF position, \mathbf{p}_{2d} is a position of SURF point in the 2D image plane.

The brief explanation of these equations is that they change the coordinate from the camera oriented coordinate into the object oriented coordinate, and then depth information is calculated based on the ellipsoidal model. The value of p_{3d_z} which is recovered in the first line of equation 1 is meaningless because it does not contain the 3D information. However, the position of x, y axis in \mathbf{P}_{3d} is valid. So, by calculating the angles in the 3D ellipsoidal model we can reconstruct the position of z axis using 2D SURF point in the last line of equation 1. Figure 2 shows the fitting result of ellipsoidal model and the matched SURF points in first and second frame.

B. Motion Vector Calculation using Non-linear Optimization

The Levenberg-Marquadt algorithm is a non-linear optimization algorithm which can provide the numerical solution to minimize the objective function.

We define the objective function to extract the motion information in the following equation.

$$\arg \min_{\mathbf{R}, \mathbf{T}} \sum \left\| \mathbf{p}'_{2d} - \mathbf{K}(\mathbf{R}\mathbf{p}_{3d} + \mathbf{T}) \right\| \quad (2)$$

where \mathbf{p}'_{2d} is SURF points in next frame, \mathbf{p}_{3d} is the reconstructed 3D points from 2D SURF point in current frame. The rotation matrix contains $\theta_x, \theta_y, \theta_z$ and translation matrix contains $\Delta x, \Delta y, \Delta z$.

In equation 2, \mathbf{p}_{3d} is the position of points in the object oriented coordinate and \mathbf{p}'_{2d} is the position of SURF points in the 2D image space. Also, we assume the camera matrix is given. Basically, equation 2 finds the rotation and translation matrix which minimize the distance between matched points in 2D space. The term, $\mathbf{K}(\mathbf{R}\mathbf{p}_{3d} + \mathbf{T})$ in equation 2 converts



Fig. 3. Reference image and candidate images; first column is reference image and the other are the sample of candidates

the point from the coordinate 3D space into 2D image space. The motion information could be extracted to adapt this objective function to Levenberg-Marquadt algorithm. In this way, the rotation and translation matrix could be updated using the previous motion information for each image frame.

C. Face tracking experiments

To verify the performance of face tracking we used 10 subjects from the Boston face database [7] which has the image sequences plus the ground truth of 3D head pose. In this database we only consider the face images with pose are within ± 15 degrees and extract the face images every third frames. In reality, when the person walks, the head pose does not move much between frames. Also, to meet the assumption of low frame rate we set up the database which is sampled at nearly 10 frame/sec.

The experimental procedure is following. First of all, corresponding SURF points were extracted and we calculated 3D points in the first frame given a motion vector. Then, we calculated the candidate motion vectors using randomly

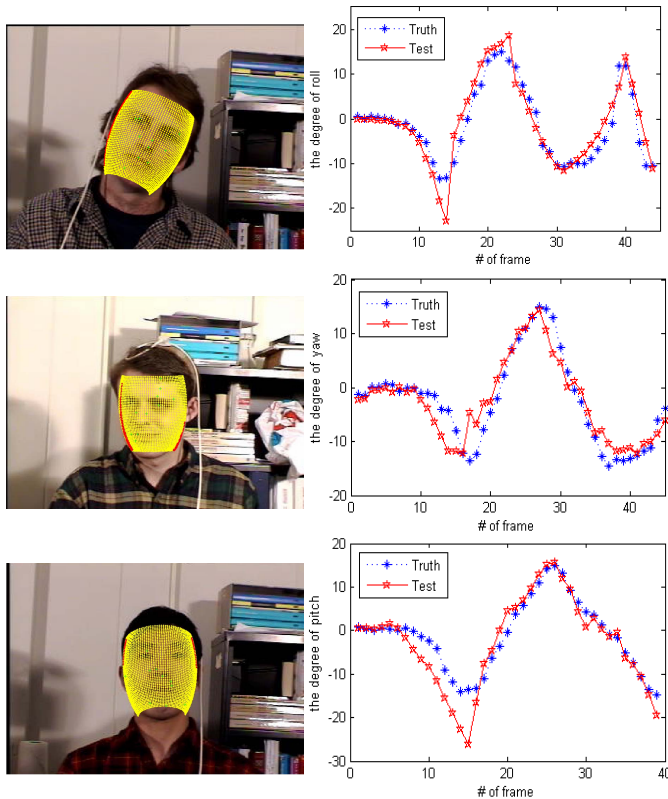


Fig. 4. The sample fitting result of roll, yaw and pitch direction

selected 3D SURF points and corresponding 2D SURF points.

From every motion vector, we constructed a 3D ellipsoidal model and extracted a 2D projection image. Then, the final motion vector which has a minimum correlation value among 2D projection images was decided. Figure 3 shows the extracted 2D projection images. Depending on a motion vector, the projection image is varied.

In figure 4, the blue dashed line represents the ground truth and the red solid line shows the test results. The average error in roll, yaw, and pitch direction are 1.79, 3.54 and 3.41 degrees, respectively, which suggest sufficient head tracking accuracy for our purpose.

III. GAIT TRAJECTORY

From previous section, the face pose can be extracted using the 3D face model under the controlled environment especially when the distance between the object and the camera is fixed. However, in reality, in visual surveillance environments there are many changes in illumination and huge head movements. Also, the low frame rate could affect the detection rate. For example, the head movement might not be continuous and the illumination could change frame by frame, especially in low frame rate video. Unfortunately, most CCTVs which are already installed in the public places have the above characteristics. To overcome the above difficulties we use not only a face image but also use alternative biometric information: gait.

A. Gait Feature Extraction

When a person is walking the movement of head must be a large and sinusoidal [21]. When a person walks in the first half of the gait cycle, the hip is in continuous extension as the trunk moves forward over the supporting limb. In the second half, once the weight has been passed onto the other limb, the hip flexes in preparation for the swing phase. As such, the specific points such as joints also show sinusoidal variation in position [22]. Further, the plane of human view varies by about ± 15 degrees when walking.

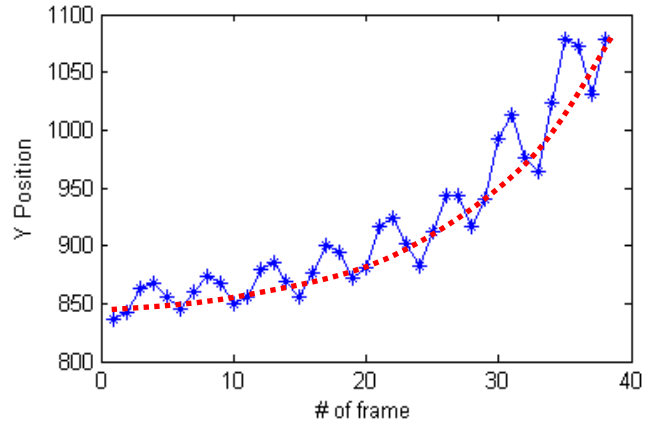


Fig. 5. A sample gait trajectory

Figure 5 represents the trajectory of the head position according to each frame. As the person walks toward the

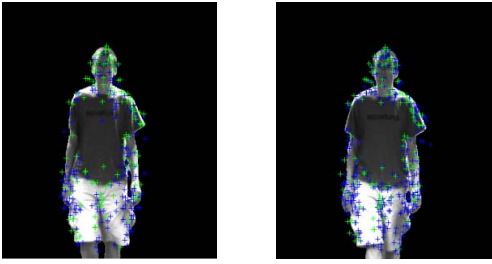


Fig. 6. SURF points of human body

camera the variance increases. It also reveals that there is periodic movement in the vertical direction. To clarify these facts we extract the exact human trajectory using some manually chosen points. We use ten samples from the biometric tunnel database [17] and extract the corresponding points for all frames. Then, we choose the centre points in human body such as the points below the neck or the points in the chest. Figure 6 shows the result of corresponding points between two images where the green points represent the corresponding points. Here, we could choose the centre point immediately below the line joining the two labeled neck points. Figure 7 show the results of horizontal and vertical trajectory respectively. As shown in these figures, the vertical gait trajectory has a consistent trend. Its nature depends on the distance between the center of the image and the position of the tracked pixel. If the position of the point is lower than the center, the trajectory increases otherwise it decreases. Unlike the vertical trajectory the horizontal gait trajectory changes with a subject's gait. Also, the variance between each point in the vertical trajectory is much larger than the variance in the horizontal trajectory. Therefore, we shall ignore the variance of the horizontal gait trajectory. In the next section we shall generate the model of the vertical gait trajectory and show the performance of model fitness using a non-linear optimization method.

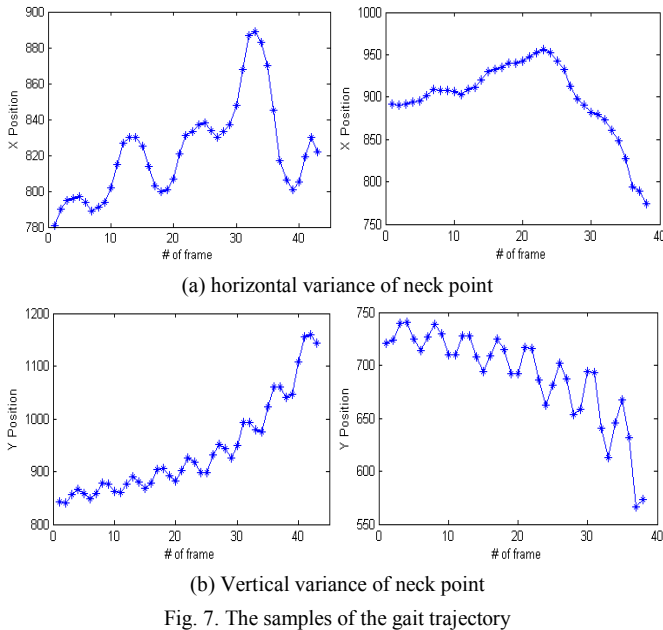


Fig. 7. The samples of the gait trajectory

B. Gait Trajectory Model Definition

To define gait trajectory model we use following assumptions.

1. The variation of z direction should be much bigger than that in the x, y directions.
2. The walking speed is constant.
3. The sampling rate is constant.
4. The testing area is not changed.

In section 3.A the gait trajectory can be divided into two parts: a periodic factor and an increasing/ decreasing factor. Under above constraints, a trajectory wave model can be defined as

$$y = \frac{K}{\alpha(1-t/\lambda)} [A_0 \sin(\omega t + \theta) + I_0] + I_{gait} \quad (3)$$

where A_0 is magnitude of initial oscillation, K is a weighting factor, ω is gait period velocity, θ is the sine wave's initial phase, I_0 is the initial bias in position, I_{gait} is initial height, λ is total walking time, and α is the length of the walking track.

To investigate the gait trajectory we can find the one period of gait trajectory has 4 to 5 gait trajectory points. So, we set the constraint of periodic variable (f) as from 1/5 to 1/3. Generally, average adult walking velocity on level surfaces is approximately 80 m/min. For men, it is about 82 m/min, and for women, about 79 m/min. Because the size of the biometric tunnel is around 6 m and the shutter speed of camera is around 100 ms the above assumptions appear reasonable. To evaluate the performance of the model we prove our model using Levenberg-Marquadt algorithm. After fitting we use R-squared, Sum of Square Error (SSE), and Root Mean Square Error (RMSE) to obtain the statistic data. Figure 8 shows the model fitting results. The blue points are the gait trajectory and the red line is the result of model fitting. We also show the autocorrelation data of fitting error. Since the gait trajectory has a periodic factor, we need to verify if the error contain the period factor. The experimental

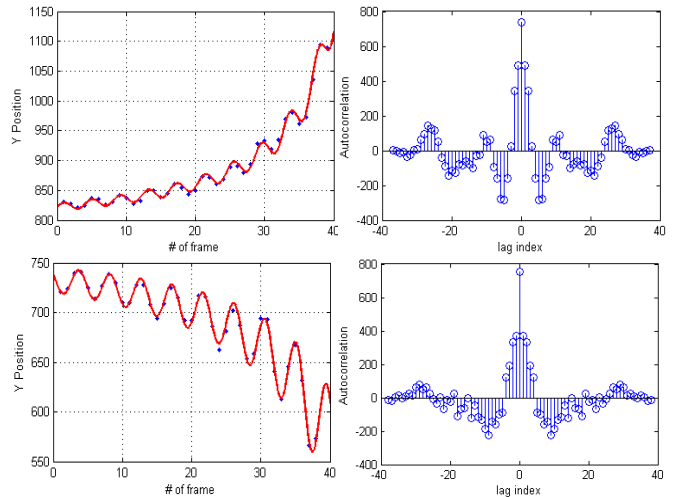


Fig. 8. The sample of the fitting result and autocorrelation of the error

TABLE I
THE NUMERAL DATA OF EACH MODEL FITTING

| | Sub1 | Sub2 | Sub3 | Sub4 | Sub5 | Sub6 | Sub7 | Sub8 | Sub9 | Sub10 |
|----------|--------|--------|--------|--------|--------|--------|-------|--------|--------|--------|
| SSE | 291.4 | 720.9 | 960.9 | 738.1 | 285.6 | 939.5 | 604.2 | 1904 | 847.1 | 753.2 |
| R-Square | 0.9983 | 0.9966 | 0.9971 | 0.9959 | 0.9979 | 0.9946 | 0.995 | 0.9932 | 0.9953 | 0.9907 |
| RMSE | 3.414 | 4.746 | 5.096 | 4.803 | 2.942 | 5.792 | 4.645 | 6.987 | 5.601 | 4.777 |

result shows that there is no specific pattern in the noise. Also, another advantage of this model is that it can express the trajectory between the trajectory points. Table I shows the numerical result of model fitting. The value of R-squared for all samples is over 99%. So, we consider that model correctly describes the gait trajectory.

C. Analysis of Gait Trajectory Model

We seek to determine a principal model for the non periodic factors in the current model.

First of all, under the pinhole camera model, the relationship in perspective projection [5] is following.

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} fX/Z \\ fY/Z \end{pmatrix} \quad (4)$$

where (x, y) is a point in 2D image plane, (X, Y, Z) are the coordinates in 3D space, and f is focal length. This is a conversion between 3D camera coordinates and 2D image coordinates. From equation 4, we can find the derivative of the perspective relationship.

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{pmatrix} = \begin{pmatrix} \frac{d}{dt} \left(\frac{fX}{Z} \right) \\ \frac{d}{dt} \left(\frac{fY}{Z} \right) \end{pmatrix} = \begin{pmatrix} \frac{\partial fX}{\partial t} / \frac{\partial Z}{\partial t} \\ \frac{\partial fY}{\partial t} / \frac{\partial Z}{\partial t} \end{pmatrix} \quad (5)$$

Here, $\frac{\partial fX}{\partial t}, \frac{\partial fY}{\partial t} \ll \frac{\partial Z}{\partial t}$ in the biometric data, and if the value of f is fixed equation 6 shows the following simple relationship.

$$\begin{pmatrix} \frac{dx}{dt} & \frac{dy}{dt} \end{pmatrix}^T \propto \left(\frac{\partial}{\partial t} \left(\frac{1}{Z} \right) \frac{\partial}{\partial t} \left(\frac{1}{Z} \right) \right)^T \quad (6)$$

When a subject walks towards the camera, the variation of x, y direction is much smaller than the variation of z direction so that the variation of x, y direction can be ignored.

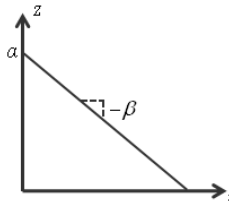


Fig. 9. The relationship between the frame number and Z direction

Assuming the velocity of walking is a constant we obtain the simple graph in figure 9. Here, Z direction is a walking

direction, α is the length of Biometric Tunnel, β is walking speed, t is time or can be thought for frame number. When the frame number increases, the distance between the camera and person decreases. Therefore, we can expand the equation in the following way.

$$Z = \alpha - \beta t \quad \alpha, \beta > 0 \quad Z, t \geq 0$$

$$\frac{\partial}{\partial t} \left(\frac{1}{Z} \right) = \frac{\partial}{\partial t} \left(\frac{1}{\alpha - \beta t} \right) \quad \text{from eq. 6}$$

$$\frac{\partial y}{\partial t} \propto \frac{\partial}{\partial t} \left(\frac{1}{\alpha - \beta t} \right) = \frac{\partial}{\partial t} \left(\frac{1}{\alpha \left(1 - \frac{\beta t}{\alpha} \right)} \right) = \frac{\partial}{\partial t} \left(\frac{1}{\alpha (1 - t/\lambda)} \right)$$

where $\lambda = \left(\frac{\alpha}{\beta} \right)$, λ is the total walking time

$$y \propto \frac{1}{\alpha(1-t/\lambda)} + a_1 \quad (7)$$

Here, a_1 is initial position.

The final equation is the same as the scaling factor of equation 3. Therefore, we can prove the gait trajectory model is correct.

IV. EXPERIMENTAL RESULTS

To automate the system, we first calculated the homogeneous matrix for each frame using the matched SURF points. Then, we chose a specific point such as the center of face or the neck position so that we can obtain a potential trajectory. After that, we apply the gait trajectory model to the potential trajectory. Since the homogeneous relationship and the trajectory between frames is known, we can extract the approximate face region of each frame to initialize the height and width of face region. In tracking, first, the 3D ellipsoidal model was applied in the first frame using a given motion vector. In the 2D projection area from a 3D ellipsoidal model we can extract SURF points so that we can remove the outliers which result in mismatching. For next frame, we track the face using a calculated motion vector in the previous frame. Also, the motion vector is calculated using the valid area mentioned above.

By this procedure, we tested the ten samples from the Biometric tunnel database. Figure 10 shows the experimental result. The first row per subject in figure 10 shows the result of extracting the approximated face regions by using the gait trajectory. We also display every sixth frame from the whole frame. As shown in figure 10, we can extract the face region regardless of pose variation, illumination change, and low resolution and the face region and pose is clearly extracted well, especially the comparison with the data from which it was extracted (figure 1).

V. DISCUSSION AND FURTHER WORK

To deploy automatic face recognition in unconstrained environments, we need to develop new techniques for

analysis. This paper describes new techniques for head pose estimation and for gait trajectory analysis. The approach to head pose estimation combines an ellipsoidal model with SURF-based feature extraction; the gait model analyses the trajectory of a looming subject. The approaches have been demonstrated with the Boston database and the Biometric tunnel database. It showed that the tracking error was around 3 degrees and the measure for gait trajectory matching was over 99%. As such gait analysis can be used to derive head trajectory when a subject walks towards the camera.

In this paper, we have a strong assumption that the walking speed is constant even though it is normal walking style of person. Therefore, there is a limitation to apply this model in some cases. Also, it is off-line system. However, this model can be applied for many facilities such as corridors, lobby, entrance of building, and so on. It is a basic research to improve the visual surveillance system. So, for the future work, we aim to translate these approaches to analyse surveillance data to provide a high resolution face image from

video data in which a subject's movement is unconstrained and the laboratory analysis is an initial step with this aim.

REFERENCES

- [1] I. Matthews and S. Baker, Active Appearance Models Revisited, *IJCV*, **60**(c), pp.135-164, 2004.
- [2] V. Blanz and T. Vetter, Face recognition based on fitting a 3D morphable model, *IEEE Trans. PAMI*, **25**(9), pp 1063-1074, 2003.
- [3] K. Liu, Y.P. Luo, G. Tei, and S.Y. Yang, Attention Recognition of Drivers Based on Head Pose Estimation, *Proc. IEEE VPPC*, 5 pp, 2008.
- [4] D.G. Lowe, Distinctive image features from scale-invariant keypoints, *IJCV*, **60**(2), pp. 91-110, 2004
- [5] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision (2nd ed.)*, Cambridge University Press, 2003.
- [6] G. D. Hager and P. N. Belhumeur, Efficient region tracking with parametric models of geometry and illumination, *IEEE Trans. PAMI*, **20**(10), pp. 1025-1039, 1998.
- [7] M. La Cascia, S. Sclaroff and V. Athitsos, Fast, reliable head tracking under varying illumination: An approach based on robust registration of texture-mapped 3D models, *IEEE Trans. PAMI*, **22**(4), 2000.
- [8] S. Basu, I. Essa and A. Pentland, Motion regularization for model-based head tracking, *Proc. ICPR*, pp. 611-616, 1996.
- [9] J. Xiao, T. Kanade and J. Cohn, Robust full-motion recovery of head by dynamic templates and re-registration techniques, *Proc. 5th IEEE Int. Conf. AFGR*, pp. 156-162, 2002.
- [10] G. Aggarwal, A. Veeraraghavan, and R. Chellappa. 3D facial pose tracking in uncalibrated videos, *Proc. PRMI*, pp. 515-520, 2005.
- [11] J.S. Jang, T. Kanade, Robust 3D Head Tracking by Online Feature Registration, *Proc. IEEE Int. Conf. AFGR*, 2008.
- [12] A. Kale, A.K. Roychowdhury, and R. Chellappa, Fusion of gait and face for human identification, *Proc. IEEE ICASSP*, pp. 901-904, 2004.
- [13] G. Shakhnarovich, L. Lee, and T. Darrell, Integrated Face and Gait Recognition From Multiple Views, *Proc. IEEE CVPR*, pp. 439-446, 2001.
- [14] X. Zhou and B. Bhanu, Integrating Face and Gait for Human Recognition at a Distance in Video, *IEEE Trans. Systems, Man, and Cybernetics*, **37**(5), pp. 1119-1137, 2007.
- [15] L. Middleton, D.K. Wagg, A.I. Bazin, J.N. Carter, and M.S. Nixon, Developing a non-intrusive biometric environment, *Proc. IEEE/R SJ IROS*, pp. 723-728, 2005.
- [16] R.D. Seely, S. Samangoei, L. Middleton, J.N. Carter, and M.S. Nixon, The University of Southampton Multi-Biometric Tunnel and introducing a novel 3D gait dataset, *Proc. IEEE 2nd BTAS*, pp. 1-6, 2008.
- [17] S. Samangoei et al., Multibiometric Datasets, In B. Bhanu and V. Govindaraju Eds: *Multibiometrics for Human Identification*, Cambridge University Press 2010 (in press).
- [18] R.M. Murray, Z Li, and S.S. Sastry, *A Mathematical introduction to robotic manipulation*, CRC Press, 1994.
- [19] K.H. An, M.J. Chung, 3D Head Tracking and Pose-Robust 2D Texture Map-Based Face Recognition using a Simple Ellipsoid Model, *Proc. IEEE/R SJ IROS*, pp. 307-312, 2008.
- [20] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, SURF: Speeded Up Robust Features, *CVIU*, **110**(3), pp. 346-359, 2008.
- [21] Aristotle, On the Motion of Animals, B.C. 350 (available at http://classics.mit.edu/Aristotle/motion_animals.html, 15/4/2004)
- [22] M. S. Nixon and J. N. Carter, Automatic Recognition by Gait, *Proceedings of IEEE*, **94**(11), pp.2013-2024, 2006.

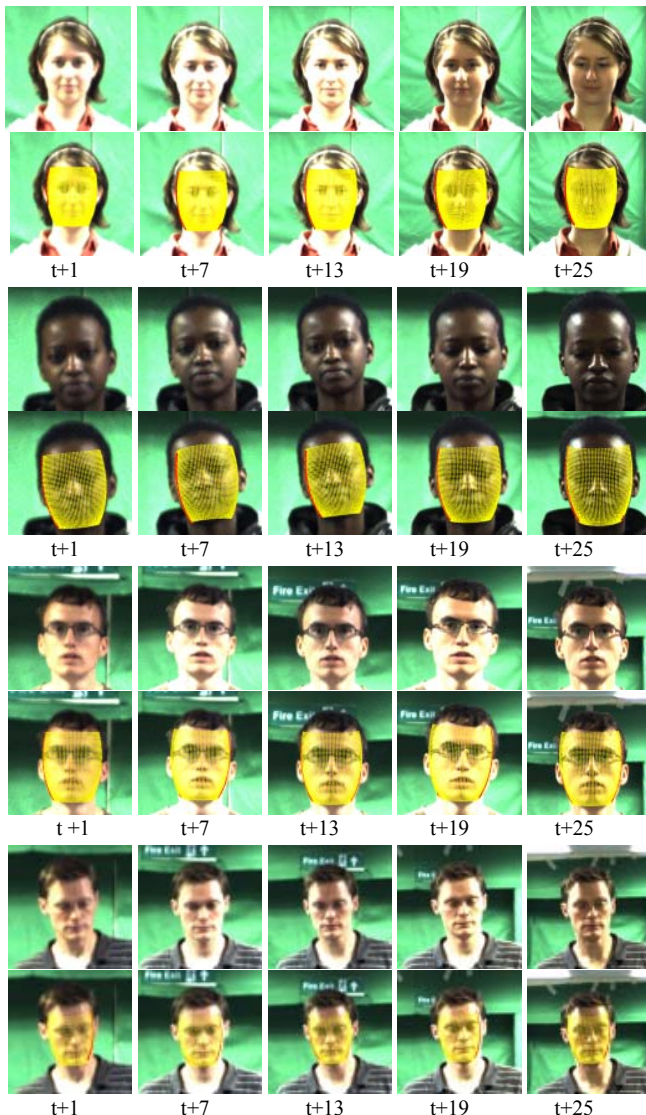


Fig. 10. The approximated face region by gait trajectory