Estimation of 3D Head Region using Gait Motion for Surveillance Video

Sung-Uk Jung and Mark S. Nixon

School of Electronics and Computer Science
University of Southampton, SO17 1BJ, UK
{suj08r, msn}@ecs.soton.ac.uk

Keywords: 3D Head position, gait period detection, heel strike detection, walking direction, gait

Abstract

Detecting and recognizing people is important in surveillance. Many detection approaches use local information, such as pattern and colour, which can lead to constraints on application such as changes in illumination, low resolution, and camera view point. In this paper we propose a novel method for estimating the 3D head region based on analysing the gait motion derived from the video provided by a single camera. Generally, when a person walks there is known head movement in the vertical direction, regardless of the walking direction. Using this characteristic the gait period is detected using wavelet decomposition and the heel strike position is calculated in 3D space. Then, a 3D gait trajectory model is constructed by non-linear optimization. We evaluate our new approach using the CAVIAR database and show that we can indeed determine the head region to good effect. The contributions of this research include the first use of detecting a face region by using human gait and which has fewer application constraints than many previous approaches.

1 Introduction

Determining the position of the head is basic step in automatic face recognition and can also be used in privacy aware surveillance (to mask the head region) or to improve person location in tracking and behaviour analysis applications. In visual surveillance there are many constraints on the ability to detect and recognize people. Depending on which CCTV technology is used, the illumination conditions, the frame rate, and the resolution can differ. Besides that, if the Region Of Interest (ROI) is considerably smaller than the captured image, conventional approaches could fail. Even though Viola's method [1] is well known for its ability to detect faces, the above constraints could lead to possible failure when detecting people in a visual surveillance environment. An alternative is to use gait, and this is immediately beneficial since the body is a larger ROI than the human face.

There are many approaches to detect a human in visual surveillance. Face region detection can be one of sub-categories of a human detection. Previous human detection methods can be classified into two main approaches. The first is a 3D assisted approach which uses view geometry and a 3D human shape model. Mohedano et al. [2] built a multi-camera geometry-based 3D tracker which uses multi-dimensional background subtraction and human template correlation. This method detected people even when they were occluded by static foreground objects. Li and Leung [3] defined Human Perspective Context according to the camera tilt angle. Then, using Model Estimation–Data Tuning the human shape and head/foot position were detected. Saboune and Laganiere [4] generated a human upper body 3D model and a likelihood function. Then, Explorative Particle Filtering was applied to detect people and for 3D tracking. In another approach, Jean et al. [5] used the 2D trajectories of both feet and the head extracted by using the silhouettes. After that, the fronto-parallel normalized view trajectory was generated from a homography transformation based on the 3D walking plane.

An alternative approach is a local feature based approach which uses an object's information such as the pattern of a face and skin colour. Li et al. [6] estimated the number of people in surveillance scenes using a Mosaic Image Difference based foreground segmentation and Histograms of Oriented Gradients for head-shoulder detection. Yang et al. [7] detected basic human actions such as placing objects and pointing using a set of motion edge history images and tree-structured boosting classifiers. Leykin and Hammound [8] tracked a subject's body and estimated the visual attention field from head pose estimation by combining a skin colour detector with the direction of motion. Chen and Chen [9] proposed a novel cascaded structure called meta-state to boost the performance of AdaBoost detection algorithm [1].

Our approach has a different starting point. We focus on detecting the head region based on the characteristics of human walking. In other words, we propose a gait-based face region detection method. First, we calculate the potential head trajectory between frames by using a homogeneous relationship. Wavelet decomposition is used to detect the component which contains a specific frequency of human walking. By analysing this component the gait cycle can be calculated. Based on the gait period and the known camera projection matrix, the heel strike position and walking direction in 3D are calculated by using our previous research [19]. After that, we define an objective function which can be used to fit the 3D gait trajectory model with the actual data by comparing a 2D potential gait trajectory with a projected 3D gait trajectory which minimizes the error of objective function. In this way, we can estimate the region of the head in 3D space.

2 Gait period estimation

The gait period is the key information to generate heel strike position and 3D trajectory correction. When people walk there is conspicuous sinusoidal head movement in the vertical plane. The highest point in a gait cycle is when both feet cross (stance) and the lowest point is when the gait stride is the largest (heel strike). Therefore, the vertical position is a cue for gait cycle detection. First, the homogeneous relationship between the adjacent images is calculated using Scale Invariant Feature Transform (SIFT) feature matching [10]. Then, the potential trajectory such as the movement of head is extracted. By wavelet decomposition different frequency signals can be analysed. These contain a signal with the same gait period as the original signal.

2.1 Homogeneous matrix calculation

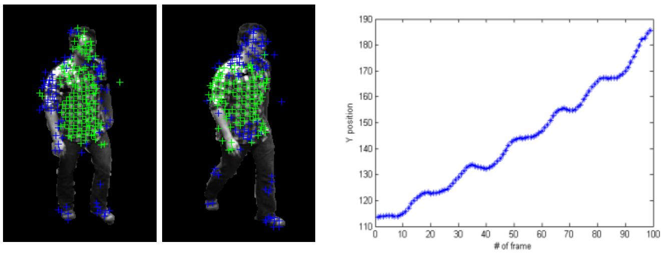
Essentially, to extract the potential trajectory, a point is determined in the first frame and its position in successive frames is estimated. In this paper, we calculate the homogeneous relationship based on analysing corresponding points detected by SIFT [10]. The main reason for adapting this method is that using all features from a human body is likely to be more robust than using local region features such as the head and legs in terms of whole movement tracking.

As a pre-processing step, the subject's silhouette image is calculated from the intensity and the colour difference (between the background image and foreground image) at each pixel [11]. Then, SIFT points are extracted from every image. From the randomly sampled eight points, each 2D homography matrix can be calculated. The 2D homography matrix describes the projective transformation between two images. The homography matrix (**H**) satisfies the following relationship [12].

$$\mathbf{H}\mathbf{x}_i = \mathbf{x}'_i \qquad (1)$$

where $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$ describes the corresponding points between two images

The Direct Linear Transformation (DLT) [12] is used to solve the 2D homography from the corresponding points. If there are many SIFT matching points between two images several homography matrixes could be generated. To find the optimized homography matrix RANdom SAmple Consensus (RANSAC) [13] is applied to choose the homography matrix which has the largest number of inliers. Figure 1 shows the result of corresponding points between two images where the green and blue points represent inliers and outliers from SIFT matching, respectively. Note that, most of the SIFT points from the upper body region are selected as inliers and those of the lower part are selected as outliers. In our experience, most of inliers are detected in the upper body because we assume that the person is walking continuously. Therefore, the lower body part such as legs and feet has more movement than the upper body. Due to this point, the homography matrix which has consistent movement between frames can be selected by RANSAC. Therefore, the homogeneous matrix can express the trajectory of the upper body.



(a) 98th frame (b) 105th frame (c) trajectory of head

Fig. 1. A sample gait trajectory

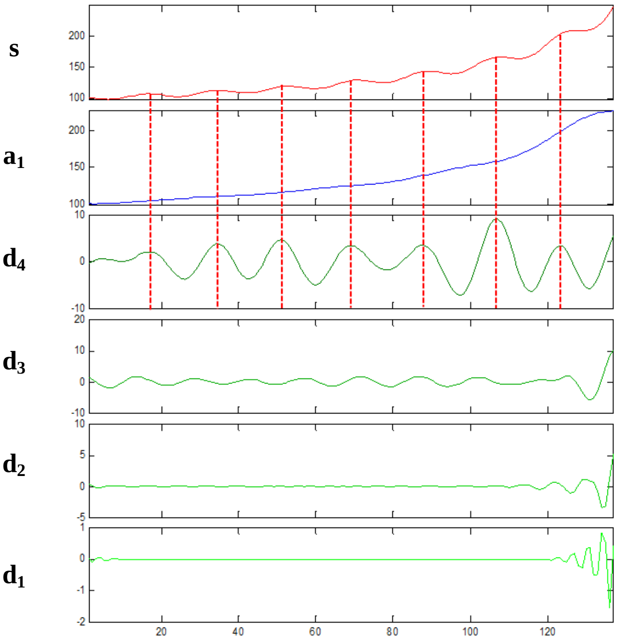
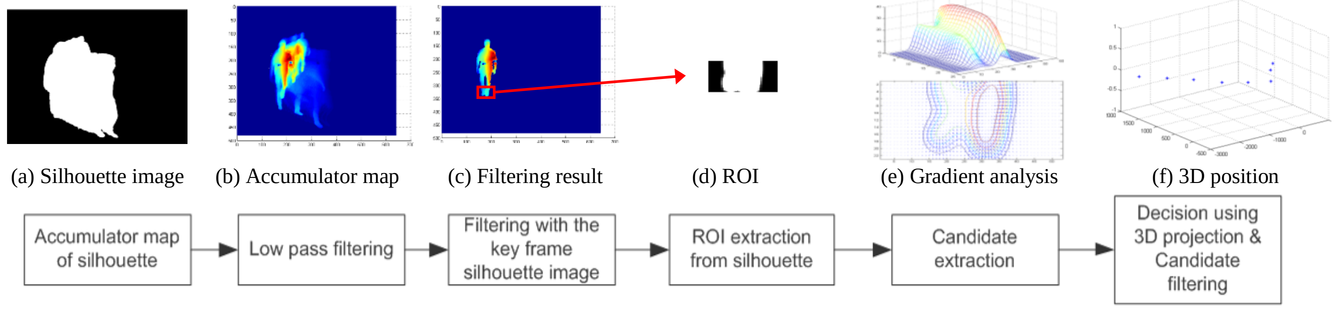


Fig. 2. Wavelet decomposition; **s** is the potential trajectory (original signal), $\mathbf{a}_1$ is the scaling coefficient, and $\mathbf{d}_1$-$\mathbf{d}_4$ are wavelet coefficients

Using the calculated homogenous matrix between frames the trajectory of a point can be extracted at each frame if the starting point is given at the first frame. The starting point could be any point on the upper body. To initiate the procedure the centre of the head is labelled manually. Figure 1(c) shows the results of the vertical trajectory. In this figure, the vertical gait trajectory has a consistent trend. The gait trajectory can be divided into two parts: a periodic factor and a scaling factor. The periodic factor is proportional to walking position; the scaling factor when a person walks toward camera. In [14], we demonstrated this looming effect when a person walks towards camera. However, given the camera projection matrix this looming effect can be removed by 3D analysis.

2.2 Wavelet decomposition

The Fast Fourier Transform (FFT) analysis can analyze the main frequencies within a signal. However, it cannot be localized in time but only in frequency. Autocorrelation is usually used to find repeating patterns, such as a periodic signal corrupted by noise. This method is also not suitable here because the gait trajectory contains a non-linear scaling factor. Given a model,


(a) Silhouette image (b) Accumulator map (c) Filtering result (d) ROI (e) Gradient analysis (f) 3D position

Accumulator map of silhouette
Low pass filtering
Filtering with the key frame silhouette image
ROI extraction from silhouette
Candidate extraction
Decision using 3D projection & Candidate filtering

Fig. 3. Process of heel strike detection

the frequency can be found using curve fitting. However, this method needs an accurate model in advance. Wavelet decomposition can be localized in both time and frequency so we use this method to detect the gait period. Because the walking period changes with speed of walking, a more specific method is needed for generalized analysis. Ladetto et al. [15] reconstructed the gait walking signal using wavelet decomposition to refine the signal. Unlike [15]'s purpose, this method is utilized to detect the gait period.

In wavelet decomposition the signal could be expressed as

$$f(x) = \sum_{k} a_{j_0}(k)\varphi_{j_0,k}(x) + \sum_{j=j_0}^{\infty}\sum_{k} d_j(k)\psi_{j,k}(x) \qquad (2)$$

where $\varphi(x)$, $\psi(x)$ are the scaling and the mother wavelet function. $a(k), d(k)$ are the scaling and the wavelet coefficient, respectively.

The 'meyer' type of the mother wavelet is used [16]. The parameters $\mathbf{a_1}$, $\mathbf{d_1}$-$\mathbf{d_4}$ are the results of decomposition of the original signal (signal **s**). As shown in figure 2, the level 4 coefficient ($\mathbf{d_4}$) has the same period as the original signal. 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. Therefore, the walking signal has a fixed range of frequency so that the signal including the walking frequency can be extracted by wavelet decomposition. In this research, the frame number is chosen as a key frame when the highest and lowest of the gait trajectory occurs. Moreover, the gait period can be detected by finding the peak position of level 4 wavelet coefficient, $\mathbf{d_4}$.

3 Heel strike detection

Heel strike detection is an important cue for human gait recognition and detection in visual surveillance since the heel strike position can be used to derive the gait periodicity, stride and step length. In our previous research [19], we outlined a method of heel strike detection based on 2D gait trajectory model which assumes that the walking speed is constant. Here, we extend the approach to be invariant to walking speed. Figure 3 shows the overall structure of the heel detection. The wavelet decomposition shows the periodicity of walking so the silhouette image and the frame number in which heel strike takes place are saved. This is the process of key frame extraction. Another step is a building accumulator map of silhouette pixels from the whole image sequence (fig. 3(b)). The silhouette image which was saved at the key frame is used to filter the accumulator map (fig. 3(c)). Then, ROI (fig. 3(d)) which is the low part of filtered accumulator map is extracted. Then, Gradient Descent is applied to obtain the maximum point which is considered as a heel strike candidates (fig. 3(e)). Finally, using the given camera projection matrix and the candidates which are calculated from other frames the final heel strike positions in 3D space are reconstructed.

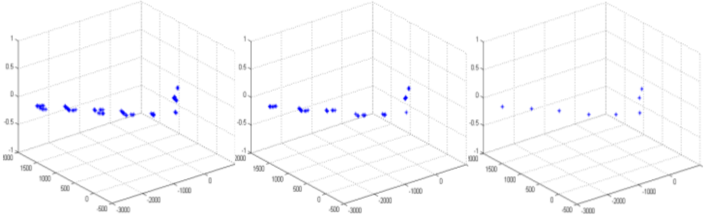
3.1 Key frame calculation

The highest point is the moment when feet cross and the lowest point is the moment when the gait stride is largest. In one gait cycle the highest point and lowest point can be calculated by detecting the maximum and minimum point from the components of Wavelet decomposition. In this way, the frame of the highest position is chosen as a key frame and the silhouette images at the key frames are saved.

3.2 Heel strike candidate extraction

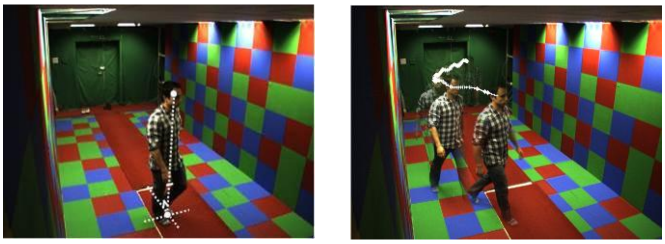
This section shows the process of detecting heel strike position using the pre-calculated key frame information. An accumulator map (which is derived by adding samples of the walking subject's silhouette from all of image sequence) is used to determine which parts of the body remained longest at same position. Generally, during the strike phase, the foot of the striking leg stays at the same position for half a gait cycle, whilst the rest of the human body moves forward. The accumulator map of a silhouette is the number of silhouette pixels. Low pass filtering is deployed to smooth the accumulator surface in fig. 3(b). Figure 3(b), (c) shows an accumulator map and filtering result of a key frame silhouette image. The colour in the figure indicates the number of silhouette pixels from blue (few) to red (many). As shown in fig. 3(b), (c) the heel strike region can clearly be distinguished from the other body parts.

The filtered accumulator map in fig. 3(c) shows the distribution of the number of silhouette pixels. It reveals that the position of heel strike has a relatively higher distribution than other regions. Using the characteristic, a Region of Interest (ROI), which is one eighth of person's silhouette height from the bottom of silhouette, is extracted (fig. 3(d)). We presume that the ROI contains the approximate heel strike region. Accordingly, the heel strike position can be extracted



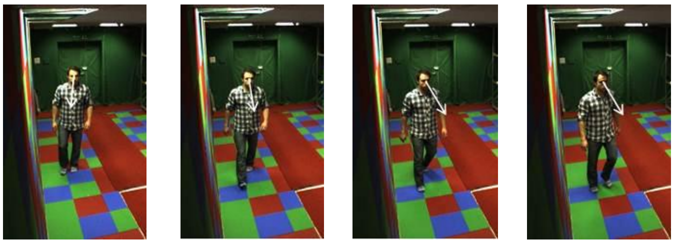
(a) The candidates (b) Filtering using other candidates (c) Filtering using 3D position

Fig. 4. The candidate filtering



(a) Position of head and heel strike (b) Head trajectory

Fig. 5. 3D head position



t+80 t+85 t+90 t+95

Fig. 6. Head position and walking direction change

by Gradient Descent. Figure 3(e) shows the three dimensional view of the extracted ROI. Figure 3(e) shows the result of analysing fig. 3(d) using the Gradient Descent. The small arrow in the figure is the point where the orientation has changed. Figure 3(e) shows the trace convergence to the local maximum.

3.3 Heel strike detection in 3D space

The key frame process also marks the frames where the feet are expected to be in the heel strike position. These estimation need to be refined to remove erroneous candidates. To reduce the invalid candidates, the key frame is calculated when the position of *y* is the lowest in one gait cycle (fig. 2) and the same procedure is executed in Section 3.2 to find other heel strike candidates. The apparent height is the lowest when the stride is maximum and both feet are in contact with the floor.

The distance between these two groups of candidates (at the highest and the lowest *y* coordinate) is calculated. Then, the candidates in the fixed distance (here, 5 pixels) are selected from the group of candidates of lowest values for *y*. As shown in fig. 4(b), after this filtering process the invalid candidates from another foot are removed. The accumulator map depends on the camera view and once the camera is calibrated the invalid candidates could be removed using back projection from a 2D image plane into a 3D world space. Using 3D projection the candidates which are the closest to the camera are selected. Since a single camera is used in our approach, we assume that a ground floor is known, i.e. z=0 (the *z* axis is vertical position). This enables calculation of the intersecting points between the projection ray from 2D image points and the ground floor. The closest heel strike to the floor is considered as the final heel strike position, thereby filtering the invalid positions. Figure 4 shows a result of the filtering process and the invalid points in fig. 4(a) are removed to give the final result in fig. 4(c).

4 Trajectory calculation in 3D space

The gait trajectory in 2D space contains a non-linear factor; the looming effect [14] which can affect the accuracy of the gait trajectory. Therefore, the domain is changed from the 2D image plane into 3D space using the camera projection matrix because in 3D space there are many advantages such that the height of walking person is constant and the looming effect can be ignored. Besides that, it can provide not only the walking direction but also the walking speed. Therefore, in this section, a 3D gait trajectory model is built based on pre-calculated 3D heel strike position and gait period. Then, using the Levenberg-Marquadt optimization method, this model is fitted to actual data to estimate the 3D position of head.

4.1 Head position while changing walking direction

Before defining the 3D gait trajectory, we need to investigate the relationship between walking direction and head position because the walking trajectory is not always the same as the trajectory of head. For instance, if a person changes the walking direction a half gait cycle gap takes place between walking trajectory and head trajectory.

Generally, the head position is located directly above the heel strike position at the moment of crossing feet. Figure 5(a) shows the sample of this relationship. In this figure, the lower white dot is the detected heel strike position and the higher white dot represents a potential center position of head. As shown in fig. 5(a), the angle between the ground floor and the line of the two points are perpendicular. Figure 5(b) presents the head trajectory from all frames. This is one of results in next section where a 3D gait trajectory model is constructed and the 3D gait trajectory is projected into 2D space. If the height of object is constant the movement of head shows the sinusoidal wave. Moreover, it can show that when feet cross the position of head is the highest and when the gait stride is the largest the position is the lowest.

Another factor to be considered is when a person changes his/ her walking direction. Figure 6 shows the moment of changing a walking direction. As shown in this figure, first, a person looks at the direction of walking. Then, a half gait cycle later the direction of feet changes. In other words, the moment that a head turns is when the gait stride is the largest. Therefore, the middle position of the 3D heel strike is used (assuming the middle position of heel strike is when a gait stride is the largest). Figure 7 presents a sample of the heel strike and middle position. The blue asterisk point represents 3D head position. The red asterisk and green triangle point

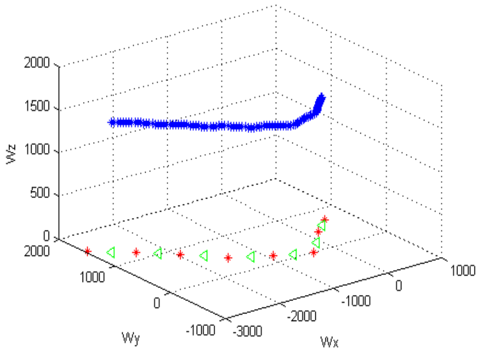


Fig. 7. Heel strike position and the middle position

are 3D heel strike position and the middle of heel strike, respectively. From the above assumptions the 3D gait trajectory model will be explained in next section.

4.2 3D gait trajectory model

Given the gait periods and 3D heel strike positions, the gait trajectory can be modelled by the series of simple sinusoidal waves which have the same magnitude and height. The potential 2D trajectory could be inaccurate. Error could be derived from the homogeneous matrix and camera projection matrix. Therefore, the gait trajectory can be corrected and the model parameter is calculated by model fitting using non-linear optimization method.

The gait trajectory model in 3D space can be defined by following relationship.

Define a walking direction vector, $\mathbf{i}_k$

$$\mathbf{i}_k = (\mathbf{H}_{k+1} - \mathbf{H}_k) / \left\| \mathbf{H}_{k+1} - \mathbf{H}_k \right\| \tag{3}$$

Then, position at ground plane and at vertical position

$$\mathbf{P}_k(t) = \mathbf{i}_k * (gait\, stride / \#of\ sampling) * t \tag{4}$$

$$z_k(t) = C_1 \sin(2\pi f_k t + \theta_k) + C_2 \tag{5}$$

Let $\mathbf{G}_{3D,k} = (\mathbf{P}_k, z_k)$ (6)

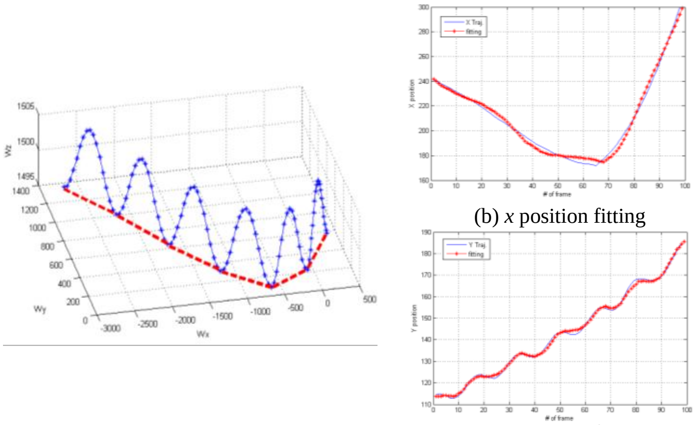
$$\mathbf{T}_{3D}(t) = \sum_{k=1}^{Num\ of\ Heel\ Strike} \mathbf{G}_{3D,k}(t) * R_k(t) \tag{7}$$

$$R_k(t) = \begin{cases} 1 & \sum_{n=1}^{k-1} p_n \le t < \sum_{n=1}^{k} p_n \\ 0 & otherwise \end{cases} \tag{8}$$

The objective function is

$$\arg\min_{C_1, C_2} \left\| \mathbf{T}_{2D}(t) - \mathbf{P} * \mathbf{T}_{3D}(t) \right\| \tag{9}$$

First, we define a 3D gait trajectory model per each period ($\mathbf{G}_{3D,k}$) which consists of x, y position ($\mathbf{P}_k(t)$) and z position ($z_k(t)$) based on the gait period (f_k) and the middle heel strike position ($\mathbf{H}_k$). Then, whole gait trajectory model ($\mathbf{T}_{3\mathrm{D}}$) is constructed by adding each model. Equation 3-8 is the details of this procedure where p_n is a period of each gait cycle and R_k is a simple rectangular function. The objective function is the equation 9 where $\mathbf{T}_{2D}(t)$ is the potential trajectory in 2D image plane (chapter 2.1), $\mathbf{T}_{3D}(t)$ is the 3D gait trajectory model, and $\mathbf{P}$ is camera projection matrix. The purpose of Eq. 9 is to calculate the magnitude (C_1) and height (C_2) which minimize the value of objective function. The objective function is an error function between 2D potential trajectory and the projected trajectory from the 3D gait trajectory model. To calculate Eq. 9 Levenberg-Marquadt algorithm is applied.



(a) 3D gait trajectory model

(b) *x* position fitting

(b) *y* position fitting

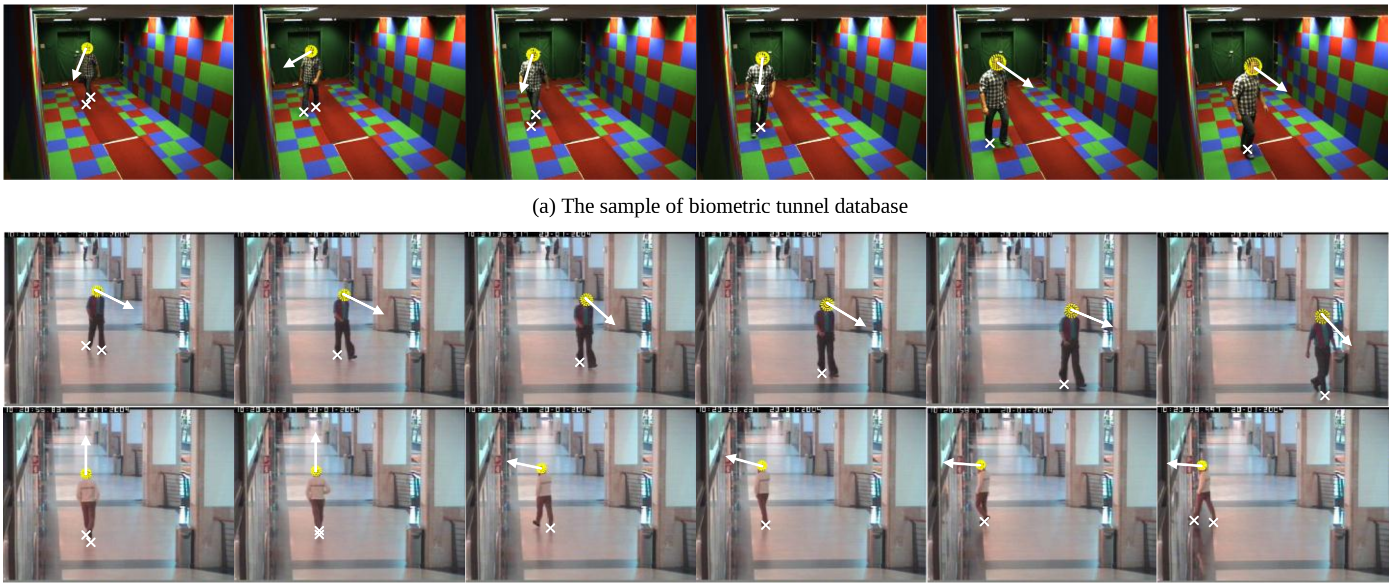
Fig. 8. 3D gait trajectory model fitting

Figure 8 displays the fitting result using the 3D gait trajectory model. Figure 8(a) shows the reconstructed 3D gait trajectory. The blue line is the trajectory of the head and the red line is the direction of head moving. Figure 8(b) and 8(c) show 2D image plane fitting after optimization procedure. The blue line is a potential trajectory of head and the red line is a fitting results.

5 Experimental results

To evaluate the proposed 3D head region estimation method we analysed the samples from CAVIAR dataset [17]. The dataset consists of 24 samples (18 males, 6 female, with around 100 images in each sequence) and are resized to 640 × 480 pixels. Each sample has a random walking direction and speed. This database is recorded at a shopping centre viewing a corridor. To calculate the camera projection matrix we estimate (perspective) corresponding points based on provided ground truth position (15 pairs of corresponding points are used).

Figure 9 shows the detection results with different environments; the biometric tunnel [18] and a shopping centre from the CAVIAR dataset [17]. The yellow region in the figure is a 2D projected face region from the 3D estimated face region. The white cross is the result of back projection from the 3D heel strike position and the white arrow represents the walking direction. As shown in fig. 9 the proposed method can estimate the head region regardless of the walking direction and speed since the method uses the basic characteristic of gait: heel strike position. Table I shows the errors between the potential 2D gait trajectory and the projected gait trajectory from 3D trajectory extracted by 3D gait trajectory model for 24 samples. We use Root Mean Square Deviation to evaluate the performance. The average



(a) The sample of biometric tunnel database

(b) The samples of CAVIAR database

Fig. 9. Detection result with different walking direction

error in *x*, *y* axis is under 4 pixels and 2 pixels, respectively. Compared to image size, this value is under 1%, suggesting good accuracy.

TABLE I
FITTING ERRORS

x pos. error	*y* pos. error	*x* err./ image width	*y* err. /image height
3.57 pixels	1.52 pixels	0.55 %	0.32%

6 Conclusions

This paper describes new techniques for head region estimation in 3D space, which are less constrained than previous approaches and can handle any direction of walk, even away from the camera. The approach to head region estimation combines 3D geometry information with human walking characteristics. In other words, from the movement of the head we can estimate the heel strike, the walking direction, and the 3D head region. The new approaches have been demonstrated with the samples from the CAVIAR database. The results show the head region estimation method is accurate even with changes in walking direction and speed. As such gait analysis can be used to derive the head region invariant to the view of camera, leading to a more versatile method of finding the human head in surveillance applications.

References

[1] P. Viola and M. Jones, "Robust real-time object detection", *IJCV*, **57**(2), pp. 137-154, Feb. 2001.

[2] R. Mohedano, C. R. del-Blanco, F. Jaurequizar, L. Salgado, and N. Garcia, "Robust 3D people tracking and positioning system in a semi-overlapped multi-camera environment", In *Proc. ICIP*, 2008, pp. 2656-2659.

[3] L. Li and M. K. H. Leung, "Unsupervised learning of human perspective context using ME-DT for efficient human detection in surveillance", In *Proc. CVPR*, 2008, 8pp.

[4] J. Saboune and R. Laganiere, "People detection and tracking using the explorative particle filtering", In *Proc. ICCVW*, 2009, pp. 1298-1305.

[5] F. Jean, A. B. Albu, and R. Bergevin, "Towards view-invariant gait modeling: computing view-normalized body part trajectories", *Pattern Recog.*, **42**(11), pp. 2936-2949, Nov. 2009.

[6] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection", In *Proc. ICPR*, 2008, 4 pp.

[7] M. Yang, F. Lv, W. Xu, K. Yu, and Y. Gong, "Human action detection by boosting efficient motion features", In *Proc. ICCVW*, 2009, pp. 522-529.

[8] A. Leykin and R. Hammoud, "Real-time estimation of human attention field in LWIR and color surveillance videos", In *Proc. CVPRW*, 2008, 6 pp.

[9] Y. T. Chen and C. S. Chen, "Fast human detection using a novel boosted cascading structure with meta stages", *IEEE Trans. IP*, **17**(8), pp. 1452-1464, 2008.

[10] D.G. Lowe, Distinctive image features from scale-invariant keypoints, *IJCV*, **60**(2), pp. 91-110, 2004.

[11] G. Cheung, T. Kanade, J. Bouquet, and M. Holler, "A real time system for robust 3d voxel reconstruction of human motions", In *Proc. CVPR*, 2000, pp. 714-720.

[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision (2nd ed.)*, Cambridge University Press, 2003.

[13] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", *Communications of the ACM*, **24**(6), pp. 381-395, 1981.

[14] S. U. Jung and M. S. Nixon, "On using gait biometrics to enhance face pose estimation", In *Proc. BTAS*, 2010, 6 pp.

[15] Q. Ladetto, V. Gabaglio, B. Merminod, P. Terrier, and Y. Schutz, "Human Walking Analysis Assisted by DGPS", *Proc. GNSS*, 2000, 6pp.

[16] Y. Meyer, *Wavelets and operators, Cambridge Studies in Advanced Mathematics*, Cambridge University Press.

[17] *The CAVIAR Test Case Scenarios: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/*.

[18] R. D. Seely, S. Samangooei, L. Middleton, J. N. Carter, and M. S. Nixon, "The university of southampton multi-biometric tunnel and introducing a novel 3D gait dataset", In *Proc. BTAS*, 2008, 6pp.

[19] S. U. Jung and M. S. Nixon, "Detection human motion with heel strikes for surveillance analysis", In *Proc, CAIP*, 2011, *LNCS* 6854, pp. 9-16.