EFFICIENT MOTION TRACKING USING GAIT ANALYSIS

Huiyu Zhou[*]*, Patrick R. Green*[†]*, and Andrew M. Wallace*[*]

[*]School of Engineering and Physical Sciences
[†]School of Mathematical and Computer Sciences
Heriot-Watt University
Edinburgh, EH14 4AS, Scotland
{*h.zhou|p.r.green|a.m.wallace@hw.ac.uk*}

ABSTRACT

For navigation and obstacle detection, it is necessary to develop robust and efficient algorithms to compute ego-motion and model the changing scene. These algorithms must cope with the high video data rate from the input sensor. In this paper, we present an approach to achieve improved motion tracking from a monocular image sequence acquired by a camera attached to a pedestrian. The human gait is modelled from the motion history of the camera, and used to predict the feature positions in successive frames. This is encoded within a *maximum a posteriori* (MAP) framework to seek fast and robust motion estimation. Experimental results show how use of the gait model can reduce the computational load by allowing longer gaps between successive frames, while retaining the robust ability to track features.

1. INTRODUCTION

We are investigating the development of mobility aids for visually impaired people. A single camera is attached to the pedestrian facing forward to observe the scene in the direction of motion and used to detect potential hazards. By establishing feature correspondences in successive frames, it is possible to recover information about the structure of the scene and the motion of the camera, which are essential for obstacle detection.

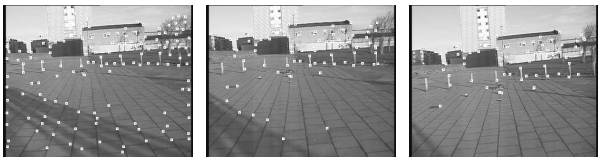
1.1. Motivation

The predict-correct strategy is adopted by many existing motion tracking systems. In the prediction stage, to avoid the effects of discontinuous motion, it is simply assumed that a minimal displacement exists between two neighboring images [1]. The potential camera motion is predicted by *linearizing* the image motion, where a tracker works reliably against a complicated model configuration. These practical techniques guarantee feature correspondence in a limited area but also incur a significant computational cost.

For tracking and obstacle detection for pedestrians, it is probable that modelling the human gait will improve the efficiency and robustness of such algorithms. For example, Molton et al. [2] modelled such motion, using a-priori data captured from a motion capture system, and in-situ data acquired from a digital compass and inclinometer. They also referred to data acquired from visual stereo and sonar devices. In this paper, our approach is to compute ego-motion only from a single video sensor during the walking sequence, i.e. not to rely on pre-training nor use additional complex devices. The advantage of this approach is that it requires minimal hardware to be attached to the pedestrian, although computational power remains an issue. Further, it does not require complex alignment and calibration, effectively learning and modifying the gait from simple video acquisition.

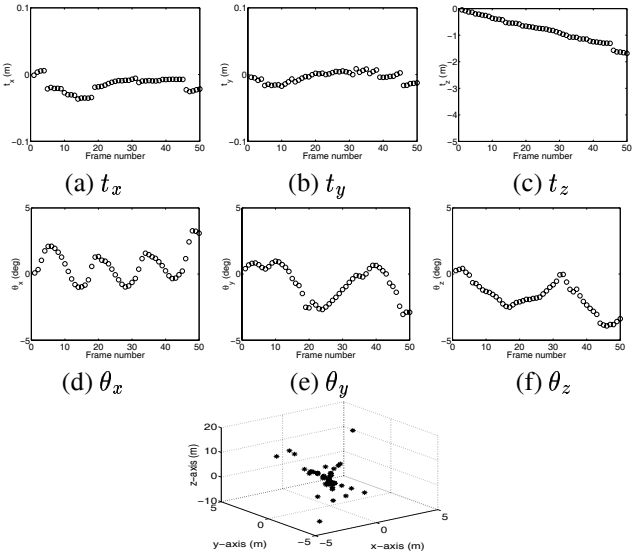
1.2. Outline of our approach

Using *structure-from-motion*, we use gait analysis to achieve approximate motion prediction, then globally refine this in the context of *maximum a posteriori* (MAP) estimation. In the first stage, the SUSAN corner detector [3] selects 150 corner features in the first image of a perspective video sequence. We then apply the Shi-Tomasi-Kanade (STK) algorithm [1] to match these features in the first two frames. Knowing the intrinsic parameters of the camera, a robust estimate of the fundamental matrix and epipolar geometry is obtained from these matched features. The camera motion between these two frames and the scene geometry are also recovered [4]. Repeating this process for n frames (n is usually 50 for a complete stride pattern at a 25 Hz sampling rate) leads to the recovery of a first estimate of a dynamic motion model with six degrees of freedom, i.e. three displacement and three Euler angles (pitch, roll and yaw). We then apply a novel predict-correct stage to continuously acquire and update the estimate of the ego-motion. The steps are: (1) Predict the feature positions in the (n+k)th frame by exploiting the 3-D feature positions and the camera mo-



(a) First frame (b) 30th frame (c) 50th frame

Fig. 1. 1st video clip with the feature points superimposed.



(g) 3-D feature positions in frame 50

Fig. 2. Estimation of camera positions and recovery of 3-D structure for the sequence shown in Fig. 1. A right-handed coordinate system is involved, where the Z-axis points in the walking direction.

tion parameters provided by the dynamic gait model. (2) Apply a coarse-to-fine strategy to match the features. (3) Apply the expectation-maximization (EM) algorithm to recover the six ego-motion parameters while removing outliers. On completion, the $(n+k+m)$th frame is examined, following the same procedure as above. The intention is to vary m, while maintaining the validity of the gait model, using all or a subset of the most recent n frames which are sufficient to sample the gait.

Fig. 1 illustrates an image sequence, and the detected and tracked feature points. Tracking these feature points leads to recovery of the fundamental matrix, camera motion and scene geometry [4]. Fig. 2 shows the camera positions based on the motion estimation, and the 3-D feature positions respectively.

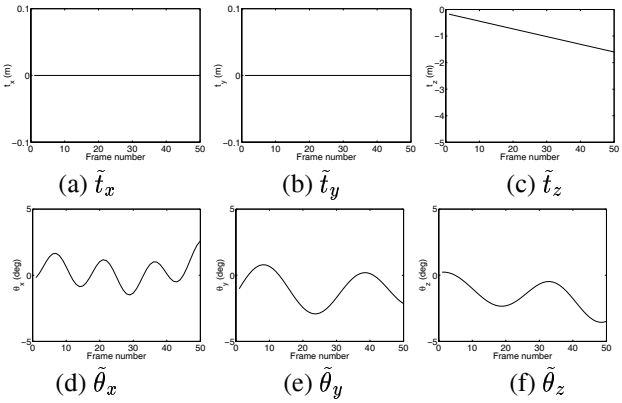


Fig. 3. Optimization of individually estimated motion parameters using IRLS.

2. GAIT EXTRACTION

Walking pattern (or gait) is highly periodic behavior that becomes mature once one starts walking at around 12 months. Knowing the ongoing gait, extracted from the motion history, we are able to predict the location of the feature points in the future frames. Therefore, tracking these features becomes more efficient.

Let d(t) be the periodic component of the generic displacement function in terms of time t during walking. The "oscillating property" of the walking pattern is described by a truncated Fourier series as

$$d(t) = d_0 + \sum_{k=1}^{N} d_t \sin(k\omega_0 + \phi_k), \qquad (1)$$

where $\omega_0 = \frac{2\pi}{T}$ (T is a time period); d_0 is the mean value; $d(\cdot)$ is one of the six degrees of freedom (DOF) in motion including rotation and translation; d_t and ϕ_k are the amplitude and phase of the k-th harmonic in one stride period, respectively [5]. The definition of N is critical, and empirical tests show 3 terms will be enough for an adequate model. Outliers usually exist in the motion estimation due to sudden changes in body posture and measurement noise. To remove such outliers and obtain a smooth model, M-estimation is used due to its robustnes and fast convergence. The iteratively reweighted least squares (IRLS) approach is employed. Fig. 3 shows the optimization results using IRLS for the motion estimation shown in Fig. 2.

3. MAP ESTIMATION

Human gait generates a 3-D "template" to predict future motion. To recover ego-motion we first register the projection of the 3-D "template" to the localized 2-D feature points. In a classical coarse-to-fine framework, the search

window for correspondence is a fixed-size square. However, in the case of a wide baseline between frames, i.e. decreasing frame rates, the search window should be simultaneously changed with the projective transformation of the images in order to guarantee efficiency and accuracy of tracking. To determine the size of the adaptive search window, the Euclidean distance (or residual) between the correctly tracked (from the STK tracking) and the predicted feature points (from the gait modelling) at the nth frame is taken into account. This frame provides the most recent information of camera motion. Experimental results show that the estimated Euclidean distribution is Gaussian or pseudo-Gaussian, with width determined by $(M + 2\sigma)$, where M is the mean of the residuals and σ is the standard deviation. As later frames are registered, these are used to determine a new adaptive search window for further tracking.

Due to correspondence errors we require robust recovery of ego-motion between frames. Consider a *dynamic* representation for the registration, which is represented as $f(\tilde{\alpha}_t, \tilde{\beta}_t, \tilde{\phi}_t)$, where $\tilde{\alpha}_t$ is the scene structure, $\tilde{\beta}_t$ is the image observation, and $\tilde{\phi}_t$ is the motion model at time t. Given a good $\tilde{\phi}_t$, $\tilde{\beta}_t$ is able to match the projected position of $\tilde{\alpha}_t$ in an image plane by maximum likelihood. Estimating motion parameters in a sequence, given the scene structure (from the previous camera motion history) and actual image correspondences, can be dealt with by the EM algorithm [6]. The conditional log-likelihood can be derived as

$$Q(\tilde{\phi}_t|\tilde{\phi}_{t-1}) = \sum_i \sum_j p(\tilde{\alpha}_{tj}|\tilde{\beta}_{ti}, \tilde{\phi}_{t-1}) \log p(\tilde{\beta}_{ti}|\tilde{\alpha}_{tj}, \tilde{\phi}_t), \tag{2}$$

where i and j label individual image features respectively.

In the E-step, the expectation of feature correspondence is based on the gait. Using [6], Eq. 2 finally becomes

$$Q(\tilde{\phi}_t|\tilde{\phi}_{t-1}) =$$

$$-\frac{1}{2} \sum_i \sum_j p(\tilde{\alpha}_{tj}|\tilde{\beta}_{ti}, \tilde{\phi}_{t-1})(\tilde{\beta}_{ti} - \gamma_{tj})(\tilde{\beta}_{ti} - \gamma_{tj})^T. \tag{3}$$

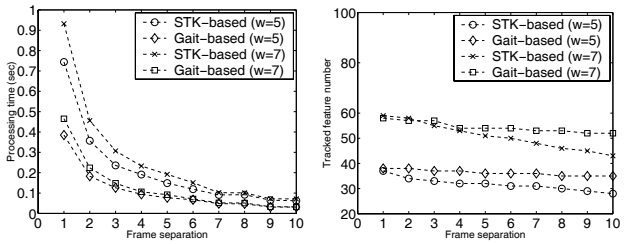
In the M-step, Eq. 3 is iterated in order to obtain the maximum likelihood. This involves finding a $\tilde{\phi}_t$ so that

$$Q(\tilde{\phi}_t|\tilde{\phi}_{t-1}) \geq Q(\tilde{\phi}_{t-1}|\tilde{\phi}_{t-1}), \tag{4}$$

where the Levenberg-Marquardt technique is applied to seek the solution for the sake of efficiency.

4. EXPERIMENTAL EVALUATION

We compare the gait-based motion tracking approach to the STK tracker, and combine this with ego-motion recovery. The intention is to show whether use of the gait model leads to improved tracking of features by a better prediction of their position in future frames, i.e. less features are lost, or alternatively whether the gait model allows us to leave a longer period between processed frames, so improving its efficiency. The STK model is used as a comparison as it is robust and widely applied.



(a) Time cost (b) Track numbers

Fig. 4. Comparison between the STK-based and gait-based framework in tracking (w - width of search window).

(a) True 55th frame (b) Texture-mapped frame

Fig. 5. Comparison of a real frame and its texture-map.

Fig. 4 summarizes a performance comparison between the STK-based and gait-based framework in tracking the sequence of Fig. 1 using a Pentium II-300 MMX PC. For fair comparison, the size of the search window is fixed. As the frame separation is increased, the processing time is reduced in an inverse relationship in each case as there is simply less processing, but the gait-based approach is significantly quicker in each case. This is due to two reasons. First, compared to the STK-based approach using a fixed number of pyramid levels, the gait-based method uses adaptive pyramid levels and subsamples, leading to less computation. Second, the STK-based motion tracking takes a long period to achieve the fundamental matrix estimation, which is not required in the gait-based method, after the feature points have been tracked. In Fig. 4 (b), the STK-based strategy loses progressively more feature points, and the difference between that and the gait-based prediction is understandably greater as the frame separation increases so that the prediction that includes the motion model is more robust. To some extent, the data flatters the STK algorithm, first because some of the features lost in each case occur simply because they leave the field of view of the camera, and second because the oscillatory motion in this se-

(a) First frame (b) 30th frame (c) 50th frame

Fig. 6. 2nd video clip with the feature points superimposed.

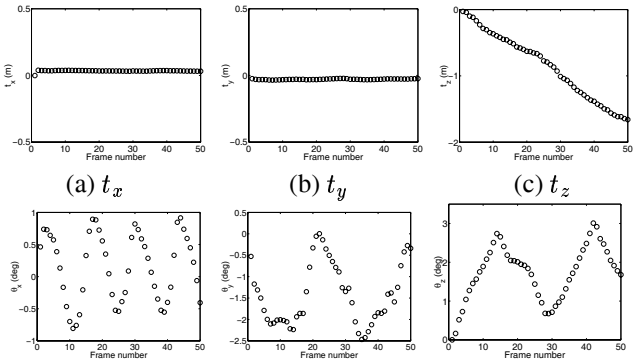


Fig. 7. Estimation of camera positions for the sequence shown in Fig. 6.

quence is not high. Fig. 5 (b) shows how, incorporating the estimated motion parameters, the recovered 3-D feature points are projected into a planar image that is then texture-mapped. This can be compared to Fig. 5 (a), showing subjectively the proposed framework performs accurately in motion estimation. Another example is shown in Fig. 6. The results are similar, but one can also observe how, as frame separation becomes larger, the time consumption of the STK-based method approaches that of the gait-based approach. In this instance, this is due primarily to the loss of the feature points during tracking, i.e. the lesser robustness of the STK algorithm results in less computational load.

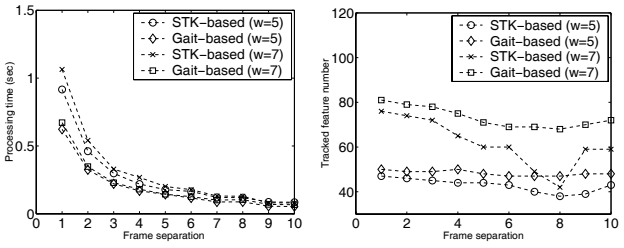


Fig. 8. Comparison between the STK-based and gait-based framework in tracking.

(a) True 5th frame (b) Texture-mapped frame

Fig. 9. Comparison of a real frame and its texture-map.

5. CONCLUSION AND FUTURE WORK

We have presented an approach to integrate gait modelling with a MAP framework to achieve faster and more robust motion tracking. The advantage of the approach is that correct modelling of the human gait results in reliable tracking of features when a greater time has elapsed between successive frames, and hence this reduces the processing burden. In addition, the more accurate prediction of feature positions in subsequent frames means that more features can be tracked successfully, hence the motion parameters can be computed from more correspondences, and the scene reconstructed with more detail. Potentially, this may lead to convenient pedestrian aids for the visually impaired as both size and power demands are reduced on wearable sensors and computers.

6. REFERENCES

[1] J. Shi and C. Tomasi, "Good features to track," in *CVPR94*, Seatle, WA, Jun 1994, pp. 593–600.

[2] N. Molton and J.M. Brady, "Modelling the motion of a sensor attached to a walking person," *Robotics and Autonomous Systems*, vol. 34, pp. 203–221, 2001.

[3] S.M. Smith and J.M. Brady, "Susan: A new approach to low-level image-processing," *International Journal of Computer Vision*, vol. 23, pp. 45–78, 1997.

[4] H. Zhou, A.M. Wallace, and P.R. Green, "A multistage filtering technique to detect hazards on the ground plane," *Pattern Recognition Letters*, vol. 24, pp. 1453–1461, 2003.

[5] A. Cappozzo, "Analysis of the linear displacement of the head and trunk during walking at different speeds," *Journal of Biomechanics*, vol. 14, pp. 411–425, 1981.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the em algorithm (with discussion)," *Journal of the Royal Statistical Society: Series B*, vol. 39, pp. 1–38, 1977.