Tracking Features on a Moving Object Using Local Image Bases

Atsuto Maki [*], Yosuke Hatanaka [†] and Takashi Matsuyama
Graduate School of Informatics, Kyoto University, Sakyo, Kyoto 606 8501, Japan

Abstract

This paper presents a new method for tracking feature points on the textureless surface of a moving object. We employ a local image basis as a descriptor of each point for dealing with intensity variances due to the relative motion of the object to the light source. In particular, we propose to adaptively update the basis for enhancing its capability as the tracking proceeds. We show that the performance of the method further improves when it is coupled with Harris feature detector at a very large integration scale.

1 Introduction

Tracking feature points on a moving object is an important basic task for estimating the motion of the object or reconstructing its 3D structure from multiview images. Feature points should be specified on the object's surface in a way that their locus are continuously identified throughout the object's motion and they are typically detected where high image gradients are observed in various directions [4, 15]. The task of tracking is then to determine the corresponding points in subsequent images by comparing the local intensity distributions. Given an image sequence with relatively small motion from frame to frame, a popular technique for tracking is to minimise the sum of squared differences of images intensities [14], usually referred to as SSD.

One of the difficulties in finding the correspondence is to deal with illumination variance which occurs due to the relative motion to the light source. The work of Jin et al. [5] effectively extended the algorithm of Shi and Tomasi [14] to take account changes in illumination by an iterative optimisation, but the results are demonstrated working for only well textured object when the camera (not the object) is subject to motion. Although one can also eliminate obvious matching failures which are caused for example by specularities as outliers [17], spurious matchings as a group will be problematic. In particular, any points can easily drift on textureless surfaces whose irradiance is mainly governed by shading.

The goal of this paper is to develop an efficient technique for feature tracking on *textureless* surfaces both in terms of detecting feature points and describing the neighbouring image intensities. Our strategy is to cope with varying effect of illumination by employing the notion of local image basis. The idea is related to the work in [3] that incorporated a general illumination model [1] into motion estimation of large image regions to compensate for the minimisation of SSD. We apply the principle of the general model to local feature points rather than to large image regions such as an entire face, as suggested in [3]. Although they are required to capture extra images of the object in a static pose under various lighting conditions in advance, we propose to generate local image bases in the course of tracking. Further, we adaptively update the basis so as to improve its ability as a descriptor as the tracking proceeds. We also report our finding in choosing appropriate feature points that are suitable for tracking on a textureless surface.

2 Tracking Overview

Let us consider an image sequence $I_j(\mathbf{x})$, with $\mathbf{x}$, the coordinates of an image point, and j, an index of the frame number. Given the time sampling frequency is sufficiently high, conventional trackers assume that small image regions are displaced but their intensities remain unchanged. The tracker's task is then to compute the displacement, $\mathbf{d}$, for a number of selected points for successive frames in the sequence. Namely, the problem has been treated to find such $\mathbf{d}$ which minimises the SSD residual in the following function (although we will introduce an alternative)

$$\epsilon = \sum_{\mathcal{W}} [(I_j(\mathbf{x}+\mathbf{d}) - \bar{I}_j) - (I_{j-1}(\mathbf{x}) - \bar{I}_{j-1})]^2 \quad (1)$$

where $\mathcal{W}$ is a small image window centred on the point for which $\mathbf{d}$ is computed. $\bar{I}$ indicates the average intensity in the region considered and the subtraction of it is

[*] Presently with Toshiba Research Europe Ltd, Cambridge Research Laboratory, UK. Email: atsuto.maki@crl.toshiba.co.uk.

[†] Presently with Sony Ericsson Mobile Communications Japan.

[1] All the images of the same Lambertian surface under different lighting conditions lie in a 3D linear subspace of the space of all possible images of the surface [13] in the absence of self-shadowing.

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

to limit the effects of intensity changes between frames. Some trackers use I_1 of the initial frame instead of I_{j-1}.

The performance of a number of popular interest points detectors and descriptors are explored in the literature for the task of matching features on weakly textured surfaces across viewpoints. For example, Maximally Stable Extremal Regions [8] is useful for wide-baseline matching. Also, Harris-affine detector [10] followed by a SIFT [7] or shape-context descriptor [2] is reported [11] to be the best combinations.

For tracking multiple points on a textureless moving object, we also need to first detect points that are suitable for tracking. We then determine the method to describe the local neighbourhood of the point so that we will replace ϵ in (1) with a new residual for more stable tracking to be possible. In this paper, we base our detector on Harris corner detector at a large derivation scale and propose to employ local images basis as a method for describing the varying appearances of the neighbourhood of the detected points.

3 Point Detector at Large Scale

For textureless surfaces the problem of finding appropriate points for tracking remains difficult although robust solutions have been proposed [4, 14, 15] for well textured objects. It is because distinctive corners tend to arise on odd-shaped part of a surface, if not at extremal boundaries or self-shadows, and they can easily change their locations according to the objects' motion. Moreover, around those points it is difficult to spatially register the local surface in approximation to a planar patch. It will thus be more desirable to detect corners at points where the gradients of intensity are moderate. For this reason, we opt for detecting points at local maxima of Harris measure [4] with a large integration scale.

The Harris measure is the second moment matrix and describes the gradient distribution in a local neighbourhood of a point $\mathbf{x}$. Harris function is

$$\det\mathbf{C} - \kappa(\mathrm{tr}\mathbf{C}) \tag{2}$$

with $\mathbf{C}(\mathbf{x}, \sigma_I, \sigma_D) =$

$$\sigma_D{}^2 G(\sigma_I) * \left[\begin{array}{cc} L_x{}^2(\mathbf{x}, \sigma_D) & L_x L_y(\mathbf{x}, \sigma_D) \\ L_x L_y(\mathbf{x}, \sigma_D) & L_y{}^2(\mathbf{x}, \sigma_D) \end{array} \right]$$

where κ is a constant, σ_I is the integration scale, σ_D the derivation scale, G the Gaussian and L the image smoothed by a Gaussian as

$$L_x(\mathbf{x}, \sigma_D) = \frac{\partial}{\partial x} G(\sigma_D) * I(\mathbf{x}) \ . \tag{3}$$

We then choose to use a large value for σ_I so that features due to moderate intensity changes can be extracted. This operation is conceptually supported by the technique of detecting scale invariant interest points [9] in that we can also select points at which a local measure is maximal over scales. The characteristics of these points is extremely suitable for tracking as it allows us to avoid detecting points at drastic shading variance due to the complex 3D surface that are actually very far from a planar patch. Figure 1 exemplifies feature points detected on a textureless surface with different values of σ_I: 1.0 (left) and 10.0 (right).

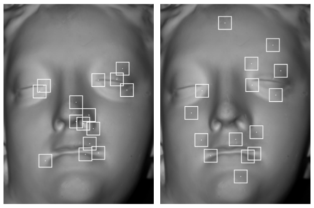

Figure 1. Local maxima of Harris measure. In the right are those by a larger integration scale, σ_I, than those in the left.

4 Local Image Basis as a Descriptor

Our focus is then to describe the local neighbourhood of the detected point in such a way that the local image characteristics can be identified for accurate tracking even under illumination variance. Invariably, a 2D image patch extracted from an image is used as the *description* of the corner feature and this description plays an important role in establishing correspondence during tracking. As is well known, however, the image patch formed about a projected corner can change dramatically over time due to the following two factors:

(a) spatial deformation from motion and projection,

(b) intensity variation from relative lighting change.

We take account both of these factors in our tracker.

Spatial Deformation

In order to accurately represent the spatial deformation of a patch, we need its local orientation, $\mathbf{n}$, and the camera projection matrices, P^j, in each frame (j is an index of the frame number). We compute $\mathbf{n}$ using tracked patches in the first few frames by deforming them while simply searching over the whole hemisphere of possible orientations to find the one that generates the most consistent intensities of pixels brought into correspondence by the particular orientation [2]. On the other hand, in the j^{th} frame, P^{j-1} is available by solving the well known structure from motion problem by factorization [16] using the coordinates of tracked points which are in correspondence until current frame.

[2] The search is quantised at every five degrees, starting with the orientation that aligns the fronto-parallel image plane.

Let us represent the spatial deformation of a local patch in j^{th} frame by a matrix, $D^j = [\mathbf{u}\,\mathbf{v}]$, using two 3-vectors, $\mathbf{u}$ and $\mathbf{v}$, which define a local plane in 3D space. For each tracked point, choosing the first frame to be canonical, the deformation in arbitrary j^{th} frame is formulated as $D^j = \mathbf{R}^j D^1$ where the elements of D^1 are available from the computed $\mathbf{n}$, and $\mathbf{R}^j$, a 3×3 rotation matrix from the first frame to the current frame, is directly given by the parameters in P^j (we assume that the fast-sampling hypothesis allows us to approximate P^j with P^{j-1}).

Dealing with Photometric Variation

For tracking a point $\mathbf{x}$, on the surface of a moving object we propose to impose a rank three constrained approximation [1] to the neighbouring pixels. For this to be possible we compute the local image basis for each feature point from a small number, a minimum of four, of initial input images. The rank three approximation is valid as long as the surface of a patch is locally illuminated by a collimated light source and somewhat three dimensional, deviating from a truly planar patch[3], so that the image basis models the changes in intensity Note that it is such points that are typically extracted at maxima of the Harris measure with a large scale.

We proceed to track feature points by initially finding correspondence of each point through first m frames $(m \geq 3)$ with a simple correlation of neighbouring image patch while considering the spatial deformation in terms of projective homography, assuming that the image variation is limited at this stage. Given registered patches, each consisting of n pixels, we record the intensities at corresponding pixels in an $n \times m$ *intensity matrix*, $\mathbf{I}_j$, each column of which contains the intensities in a single patch where j is the index to the current frame. We generate the initial estimate of local image basis, $\mathbf{B}_j|_{j=m}$, by a rank three approximation to the matrix $\mathbf{I}_j$ such that $\mathbf{I}_j = \mathbf{B}_j \mathbf{S}_j$ where $\mathbf{B}_j = [\mathbf{b}_1\ \mathbf{b}_2\ \mathbf{b}_3]$ is a 3D local image basis, and $\mathbf{b}_{1,2,3}$ span the local illumination subspace. $\mathbf{S}_j$ is a $3 \times m$ matrix whose columns correspond to collimated relative light sources up to an ambiguity, but not used in the rest of the paper.

Although the principle of using local image bases has been first introduced in [18] to compose a 3D structure model by a group of bases, each basis computed at the initial stage may naturally have limited capability as a descriptor by itself since only small variance of illumination can be encoded in the representation. Our key advance is that we propose to automatically update the linear image basis as we observe more variation of the neighbouring irradiance as the tracking proceeds.

5 Tracking Algorithm

We track each feature point by searching for a corresponding point while investigating the consistency to the linear image basis. In the j^{th} frame $(j > m)$ we deform each component of $\mathbf{B}_{j-1}$ using D^j, and solve the problem of finding such $\mathbf{d}$ that minimises the rank three residual by replacing the function in (1) with

$$\epsilon_b = \sum_{\mathcal{W}} [I_j(\mathbf{x}+\mathbf{d}) - \hat{I}]^2 \tag{4}$$

where $\hat{I}$, an estimated intensity, is the element of $\hat{\mathbf{L}}_j$ such that

$$\hat{\mathbf{L}}_j \quad = \quad \mathbf{B}_{j-1}(\mathbf{B}_{j-1}^\top \mathbf{B}_{j-1})^{-1}\mathbf{B}_{j-1}^\top \mathbf{L}_j \tag{5}$$

and $\mathbf{L}_j \in \mathcal{R}^n$ represents a local patch, containing the values of $I_j(\mathbf{x}+\mathbf{d})$, which is also determined with D^j.

Updating the Local Image Basis

After finding the corresponding point in a new frame, we update the local basis, $\mathbf{B}_j$, by additionally using the intensities at the neighbouring pixels in the patch if they are judged to be useful. That is, at each tracked point we revise the intensity matrix $\mathbf{I}_j$ by incorporating a new column, consisting of the intensities in the referred patch, and check their feasibility by the rank of $\mathbf{I}_j$ with the *evaluation ratio*, $r(j)$, of the third and the fourth singular values: $r(j) = \sigma_3/\sigma_4$. If rank$(\mathbf{I}_j)$ is three, which is the ideal case, σ_3 should be much larger than σ_4 which should always be very small. In the j^{th} frame, thus, if the value of $r(j)$ is higher than in the previous frame, we replace $\mathbf{B}_{j-1}$ with a new one, $\mathbf{B}_j$, computed by decomposing $\mathbf{I}_j$. We then discard a redundant column at the same time. This updating process enforces the basis to be more tolerant to varying effect of lighting as the tracking proceeds even if the initial basis happens to be rank deficient due to degenerate motions.

6 Experiments

Figure 2 shows the performance of our tracker for the input sequence of a moving mat statue of "Venus". The surface is textureless, which is problematic not only for SSD based trackers such as by (1) but for descriptors dependent on texture. The tracked points are indicated as the centre of white four-sided shapes which also show how the patches are registered. They are shown at every fourth frame, starting from the eighth frame (points in the first frame in Figure 1 (right)). The proposed tracker continues to track the features more accurately than does zero mean SSD with which many points are prone to drift, e.g. those around the left eye. We have tested our tracker with other sequences of texureless objects and found it to perform stably.

[3] This is in contrast to the work in [5] which assumes normal vectors do not change within the local patch.

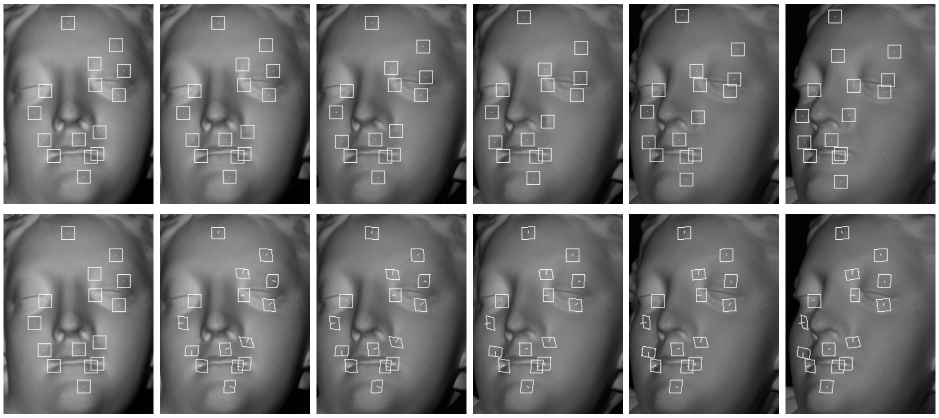

Figure 2. Tracking results for “Venus” at every fourth frame. Top: Zero mean SSD. Bottom: Proposed method.

Figure 3 shows an example of local linear image basis which is computed and updated on the right end of the mouth of “Venus”. The deformed local patches, $\mathbf{L}_j|_{j=1,\ldots,4}$, and the initial basis, $\mathbf{b}_{1,2,3}$ are in the left. The basis is shown in descending order of the corresponding singular values, from the left to the right, and the fourth (rightmost) is thus the residual. $\mathbf{L}_j|_{j=1,\ldots,4}$ include only small intensity variations between the frames and the computed basis of lower order turns out to be noisy; the third base, $\mathbf{b}_3$, looks like residual as the fourth one. In the right are new input patches, $\mathbf{L}_j|_{j=1,12,15,16}$, in which more variation of illumination is involved. The updated basis reflects 3D aspect of the surface more effectively as obvious in $\mathbf{b}_2$ and $\mathbf{b}_3$. Figure 4 shows how the evaluation ratio, $r(j)$, is updated as the tracking proceeds. $r(j)$ is initially about 2.0 but increases especially in several frames after the initialisation, indicating that the ability of the basis is enhanced to account for the illumination variances.

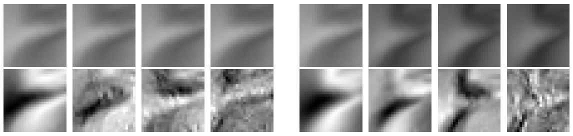

Figure 3. Left: Patches tracked in the first frames (top). Computed basis (bottom). Right: New inputs and updated basis.

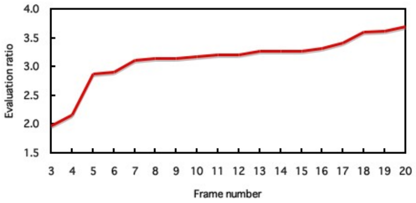

Figure 4. The evaluation ratio of a local image basis, $r(j) = \sigma_3/\sigma_4$, in each frame.

To evaluate the performance of the tracking quantitatively, we computed average distances of tracked points to epipolar lines. We first computed the affine fundamental matrices by using the coordinates of tracked points in each frame up to the 20th frame. We then drew epipolar lines in the first frame by using them and checked the average distance. Since the distance should become zero when perfect matchings were available, we can employ this value as a measure of tracking accuracy.

Figure 5 shows the average distances plotted for each frame (initial frame is numbered zero), for each case of using linear image bases and zero mean SSD. For comparisons, the results are shown for two different sets of feature points detected with different values of σ_I: 1.0 and 10.0 (See Figure 1). In the graph ’H’ and ’L’ stands for the two cases, respectively. We can observe that the smallest error (distance) is achieved in the case of L-Basis, i.e. when the points detected at $\sigma_I = 10.0$ are tracked by using the linear image basis. The error increases more drastically when using zero mean SSD regardless of the choice of the point set.

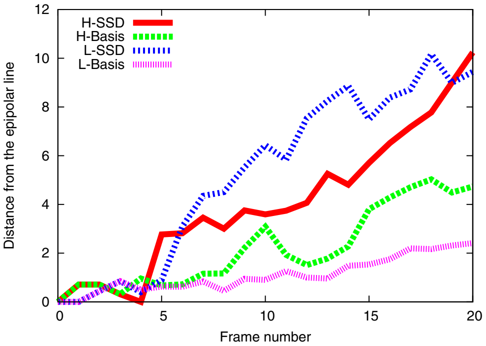

Figure 5. Average distances (in pixels) of tracked points from epipolar lines.

7 Summary

We have tackled the ill-posed problem of tracking feature points on a textureless surface and shown promising results by (i) employing a local image basis as a descriptor of each feature point which we (ii) detect by Harris measure at a very large scale. In particular, we proposed to (iii) update the bases as the tracking proceeds so that it can accommodate varying effect of illumination. Future work will be directed to evaluate the algorithm using more data, e.g. involving specularities, in comparison to other related approaches [12], and to select the optimal derivation scale [6] for feature detection by automatic scale selection.

Acknowledgment

This work is supported by MEXT, Japan, under a Grant-in-Aid for Scientific Research (No.18651077), in part by ditto (No.18049046) and by national project on ”Development of high fidelity digitization software for large-scale and intangible cultural assets”. The implementation of Harris operator available at http://www.img.tutkie.tut.ac.jp/programs/index.html was used in the experiments. We also wish to thank Lyndon Hill for his helpful comments on the manuscript.

References

[1] P. Belhumeur and D. Kriegman. What is the set of images of an object under all possible lighting conditions? In *CVPR*, pages 270–277, 1996.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE-PAMI*, 24(4):509–522, 2002.

[3] G. D. Hager and P. N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE-PAMI*, 20:10:1025–1039, 1998.

[4] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. Fourth Alvey Vision Conference*, pages 147–151, 1988.

[5] H. Jin, P. Favaro, and S. Soatto. Real-time feature tracking and outlier rejection with changes in illumination. In *ICCV*, pages 684–689, 2001.

[6] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30:2:79–116, 1998.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:2:91–110, 2004.

[8] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002.

[9] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *ICCV*, pages 525–531, 2001.

[10] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *ECCV (1)*, pages 128–142, 2002.

[11] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3d objects. In *ICCV*, pages 800–807, 2005.

[12] M. Nishino, A. Maki, and T. Matsuyama. Phase-based feature matching under illumination variances. In *IEA/AIE*, pages 94–104, 2007.

[13] A. Shashua. *Geometry and photometry in 3D visual recognition*. PhD thesis, Dept. Brain and Cognitive Science, MIT, 1992.

[14] J. Shi and C. Tomasi. Good features to track. In *CVPR*, pages 593–600, 1994.

[15] S. M. Smith and J. M. Brady. SUSAN – A new approach to low level image processing. Technical Report TR95SMS1c, Chertsey, Surrey, UK, 1995.

[16] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *IJCV*, 9:2:137–154, 1992.

[17] T. Tommasini, A. Fusiello, E. Trucco, and V. Roberto. Making good features track better. In *CVPR*, pages 178–183, 1998.

[18] C. Wiles, A. Maki, and N. Matsuda. Hyper-patches for 3D model acquisition and tracking. *IEEE-PAMI*, 23:12:1391–1403, 2001.