Target Tracking Using Mean-Shift And Affine Structure

Chuan Zhao, Andrew Knight and Ian Reid
Department of Engineering Science, University of Oxford, Oxford, UK
{*zhao, ian*}*@robots.ox.ac.uk*

Abstract

In this paper, we present a new approach for tracking targets with their size and shape time-varying, based on a combination of mean-shift and affine structure. Although the well-known mean-shift colour-based tracking algorithm is an effective tracking tool, difficulties arise when it is applied to track a size-changing visual target due to the fixed kernel-bandwidth. To improve this, the present study employs a corner detector on the object candidate from mean-shift and reconstructs the target position and relative scale between frames using the affine structure available from two or three views. In comparison experiments against previous algorithms, the present model shows better tracking-consistency and good efficiency. Our algorithm is also demonstrated in a real-time implement controlling the pan-tilt-zoom parameters of an active camera. The results indicate the model's tracking capability in the presence of scale change and partial occlusions.

1. Introduction

The efficient tracking of moving, non-rigid objects through images is one of the most important disciplines in the field of computer vision and artificial intelligence. Nearly all real-time applications, such as traffic control and automated visual surveillance require robust and accurate tracking. We are particularly motivated by demands for automated surveillance in which it is common to have significant changes in the size of a target, due to changes of its distance to the camera or simply changes of zoom.

The Mean-Shift (MS) based tracking algorithm [3] is a popular method for real-time target tracking, because it is fast, simple and its non-parametric colour-based appearance model confers a large degree of viewpoint invariance. It uses a density gradient estimator iteratively to compute the local maximum achieving the most similarity to the sample distribution. In MS tracking, the kernel scale is a crucial parameter. Not only does it directly determine the size of the window within which sample weights are examined, but also should be proportional to the expected image area of the target. Normally, the kernel scale is initialised by the first tracking window then fixed in the whole tracking process. However, when the scale of the target changes significantly, or in the worst case of the target scale exceeding the tracking window, tracking failure usually results.

To address the scale adaptation problem Comaniciu *et al.*[3] suggested a simple scale adaptation scheme which modifies the radius of the kernel profile by a certain fraction ($\pm 10\%$). In the current time step, MS is run independently for three different kernel scales and the radius yielding the largest Bhattacharyya coefficient is chosen. However when the target size exceeds the search window size, the Bhattacharyya coefficient will not force the kernel size to grow due to its characteristic of always converging to the local maximum value in a smaller search window. In [1] Bradski proposed the CAMSHIFT method: after MS converges to the new location, an ellipse is computed based on second order central moments of skin probability image pixels inside the search window slightly larger than the MS window size, then scale is updated according to the magnitude of the second moments. This procedure runs iteratively until convergence. Collins [2] proposed an adaptive mechanism for the varying MS kernel size in the scale space based on Lindeberg theory [6]. While it significantly outperforms [3], it still exhibits some problems related to growing targets, with tendency to underestimate the true scale.

To overcome this limitation of the MS method, we have developed a hybrid tracking model based on the combination of MS and affine structure of a set of feature point correspondences. Our approach is closest in spirit to the work proposed by McLauchlan and Malik in [7] which reconstructs the affine structure combined with stereo cues whilst ours combines with MS. We use the feature points analysis to recover affine parameters and estimate the relative scale between frames

from them. Additionally, the target location is updated through affine structure to avoid problems occurred in MS when objects and background are similar in colour distribution. We compare our results to the standard MS with the adaption scheme of checking $\pm 10\%$ scale change at each frame [3], and also to Collins' [2] more rigorous scale-space search.

2. PROPOSED APPROACH

We assume that the object in view is undergoing rigid (or affine) motion and the images in successive views are formed by affine projection. This latter assumption is reasonable in the context of surveillance applications in which target relief is typically small in relation to depth. We then proceed as follows: First in each time step, MS is applied to compute the inter-frame translation. This yields a candidate target location with previous kernel size to get **Window1** in the current frame. We detect corner features inside this MS window then seek corner matches with the features from the most recent frames. If more than three matches are found, affine structure is estimated from point correspondences. The relative scale of the target is then recovered from the affine transformation. We also transfer the location of the target in the previous frame to a new location (**Window 2)** in the current frame based on the affine structure using the method described by [8]. **Window** 1 and 2 are both re-sized by the relative scale. Finally, the new tracking window is determined by choosing whichever window yields the greater Bhattacharyya coefficient. In our present framework, the tracking window is supposed to be T_i centred on c_i in frame i. In frame $i+1$ the first step is applying MS to compute the new location of the tracking window $\hat{T}_{i+1}$ centred on $\hat{c}_{i+1}$. Except this step, We expand each of remaining steps in the following subsections.

2.1 Feature-Based Matching & Tracking

Reid and Murray [8] developed a method that detects a cluster of corner features on the target and tracks those corners individually over time. By computing implicit affine structure of the point cloud, it can track targets smoothly despite the inherent unreliability of an individual corner track. Furthermore, other work based on point features has shown how the relative scale, translation (and indeed other parameters such as rotation and shear) can be recovered from the affine projection matrices or affine epipolar geometry [8] [10] [11]. Though this method produces good results when successful, it is inclined to be brittle. Thus in some respects, it can be considered complementary to colour-histogram based tracking. It is this observation we leverage to produce our hybrid tracker.

Affine structure can be computed from as few as four points in three views. In practice, the differential quantities (change in translation and change in scale) we are primarily interested in can be more reliably extracted via 2D estimation (planar target assumption), for which we require as few as three points matched between a pair of successive frames. We proceed by first detecting corner features [9] in the candidate window $\hat{T}_{i+1}$ updated through MS. Subsequently, we use Zero-mean Normalised Cross Correlation (ZNCC) to seek the best matches between corner features from most recent frame and current frame. Given $n \geq 3$ such matches in two or more views, it is possible to determine the affine structure of the n point configuration. Given our previous assumptions, the point locations in each view will be related by the affine equation

$$\Delta \mathbf{x}_{i+1} = \mathbf{A}_{i+1} \Delta \mathbf{x}_i + \mathbf{b}_{i+1} \tag{1}$$

where $\Delta \mathbf{x}_i$ and $\Delta \mathbf{x}_{i+1}$ are the registered corner locations in the previous and current frames respectively, $\mathbf{A}_{i+1}$ is a 2x2 matrix and $\mathbf{b}_{i+1}$ is a 2x1 vector, together defining the 6-degree of freedom transformation between the views. Our aim at this stage is to compute $\mathbf{A}_{i+1}$ and $\mathbf{b}_{i+1}$, as well as reject outliers and this is done robustly from the set of putative point matches using RANSAC [4].

2.2 Recovering Relative Scale

As proposed by Tordoff [11], based on Shapiro's analysis [10], we use the epipolar geometry in two views to recover the relative scale of a target. More specifically, [10] describes how the spacing of the parallel epipolar lines under affine projection are characteristic of the scale, once the images are corrected for known aspect ratio. The affine epipolar constraint equation can be expressed in the form of a fundamental matrix $\mathbf{F}_A$

$$\mathbf{x}_{i+1}^{\mathrm{T}} \mathbf{F}_A \mathbf{x}_i = \begin{bmatrix} x_{i+1} & y_{i+1} & 1 \end{bmatrix} \begin{pmatrix} 0 & 0 & a \\ 0 & 0 & b \\ c & d & e \end{pmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = 0$$

where $\mathbf{x}_{i+1}$ and $\mathbf{x}_i$ are homogeneous 3-vectors representing points in the image planes, in frame $i+1$ and frame i respectively. If the image points are registered, $e = 0$. Then

$$a\Delta x_{i+1} + b\Delta y_{i+1} + c\Delta x_i + d\Delta y_i = 0$$

where $(\Delta x_{i+1}, \Delta y_{i+1})$ and $(\Delta x_i, \Delta y_i)$ are registered points in image planes. Substituting the Koenderink and van Doorn (KvD) expression in [5] for a rotation matrix into this equation, we can derive how scale is calculated:

$$s^2 = \frac{c^2 + d^2}{a^2 + b^2}$$

The values for $a \ldots d$ in the affine fundamental matrix $\mathbf{F}_A$ can either be computed from affine epipolar geometry algorithm using two views of at least four non-coplanar points (*i.e.* valid even if the planar assumption is violated), or directly from the affine transfer Eq 1:

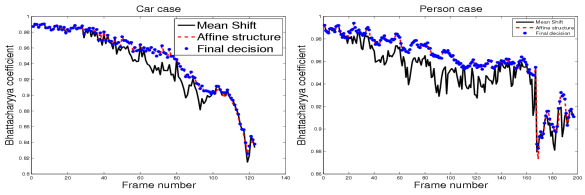
$$\begin{aligned} a &= \mathbf{A}_{[2,3]}, \quad b = -\mathbf{A}_{[1,3]} \\ c &= \mathbf{A}_{[1,3]} * \mathbf{A}_{[2,1]} - \mathbf{A}_{[2,3]} * \mathbf{A}_{[1,1]} \\ d &= \mathbf{A}_{[1,3]} * \mathbf{A}_{[2,2]} - \mathbf{A}_{[2,3]} * \mathbf{A}_{[1,2]} \end{aligned} \tag{2}$$

2.3 Post Correction

As in most colour-based tracking algorithms, the resulting location in MS will drift when the target is not well-defined (*i.e.* looks similar to the background). We make use of the affine structure to ameliorate this problem, as proposed by Reid & Murray in [8]. The registered center of the target $\Delta\mathbf{c}_i$ in the previous frame is projected by the affinity to the current frame as the new center of the target $\Delta\tilde{\mathbf{c}}_{i+1}$.

$$\Delta\tilde{\mathbf{c}}_{i+1} = \mathbf{A}_{i+1}\Delta\mathbf{c}_i + \mathbf{b}_{i+1}$$

From $\Delta\tilde{\mathbf{c}}_{i+1}$, we generate a new window $\tilde{T}_{i+1}$. After updating the size of candidate **Window1** ($\hat{T}_{i+1}$) from MS and **Window2** ($\tilde{T}_{i+1}$) from the affinity, through the relative scale computed in the previous step, we estimate the Bhattacharyya coefficient for both, select the one with a higher Bhattacharyya coefficient to update the final target window. In this way, not only the drift problem in MS is fixed, but also false matches in affine transfer are avoided. Thus we substantially reduce the risk of the drift from the true location for the fixation point and the resulting error accumulation.



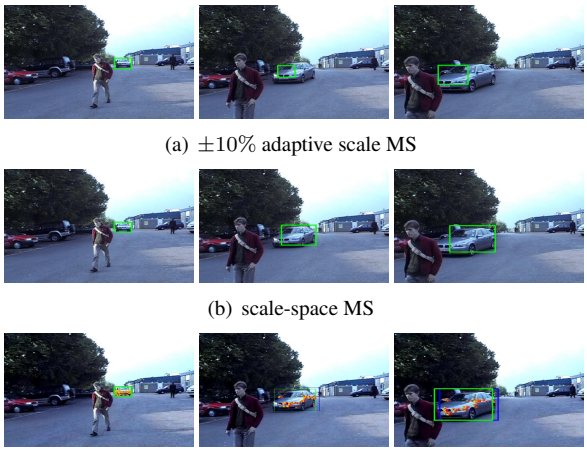
(a) car sequence 1 (b) person sequence 1

Figure 1. Confidence curves for MS-Affine tracker. solid line: the confidence of pure Mean-Shift; dashed line: confidence of fixation point from affine transfer; dotted line: confidence of final choice

Figure 1 shows the Bhattacharyya coefficient (Confidence) curves for a car sequence and a human sequence respectively. In both sub-figures, solid lines represent the confidence of pure Mean-Shift, dashed lines represent the confidence of fixation point from affine transfer, and dotted lines show the final position that is used to update. All curves drop down dramatically in last several frames because the appearances and scales of targets have changed greatly. This can be avoid if an adaptive model update scheme is applied. For nearly all numerical tests, the location recovered from affine structure yields a higher Bhattacharyya coefficient than that from MS. It confirms that we get better scale than MS and scale-space tracking. The benefit of MS tracking is that we can continue to track if there are few features, *e.g.* because of image blur, and it helps localise objects so that corner matching is easier.

3. EXPERIMENTAL RESULTS

In this section, the hybrid model is used to track targets which are undergoing scale change. We have conducted various of experiments involving vehicles and humans, outdoor and indoor. We initialise the tracking window by hand, which could be replaced by a detector (background subtraction, or AdaBoost detection) in the future. The target histogram has been derived in the RGB space with 32x32x32 bins on the image with a resolution of 640 by 480. Figure 2 and 3 show how three



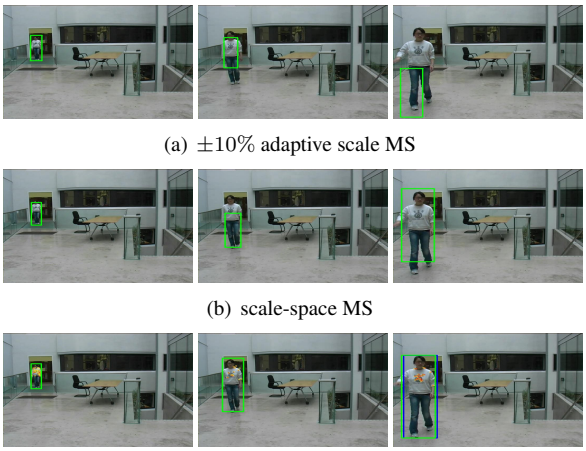
(a) $\pm 10\%$ adaptive scale MS

(b) scale-space MS

(c) MS with affine structure

Figure 2. Tracking results of three methods for the car sequence in Frame 2, 42, 112

algorithms work on "looming" targets. The images in Figure 2 are Frame 2, 42 and 112 from a sequence involving a car driving towards the camera, while Figure 3 are sample images from Frame 2, 123 and 170 in a sequence of a person walking around a test trajectory. The size of the targets is increasing through the sequence. In Figure 2(a) and 3(a), which are illustrating pure MS with $\pm 10\%$ scale adaptation, the tracking window shrinks even though the actual size of the target is increasing. In Figure 2(b) and 3(b), Collins' scale-space model-tracking has a much better performance, but it



(a) $\pm 10\%$ adaptive scale MS

(b) scale-space MS

(c) MS with affine structure

Figure 3. Tracking results of three methods for the person sequence in Frame 2, 123, 170

still sometimes tends to underestimate the scale due to target-background similarity. As a result, the MS kernel is smaller than the size of the target, which makes the tracking window roam around on a likelihood plateau around the true window of the target. The output of the tracker is unstable throughout the whole sequence. In Figure 2(c) and 3(c), the dark rectangle is the middle result computed from MS step, and the light one is the final result with relative scale and location correction from affine Transfer. The targets are consistently tracked both in location and scale. The overall performance of our hybrid model is superior to its predecessors. In the last frame of both sequences, there are fewer feature point matches than before. This is related to the significant change of the target shape and scale. This reduction in the feature point matches did not jeopardise the stability and the good performance of the model but could cause possible inaccuracy in scaling. Figure 4

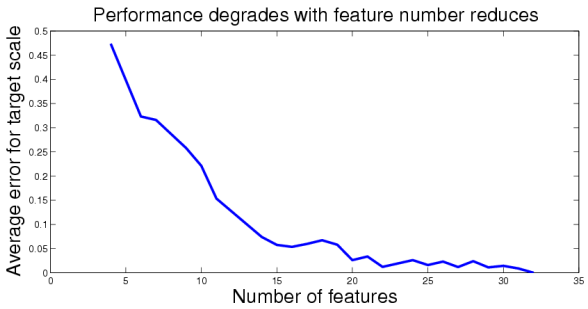
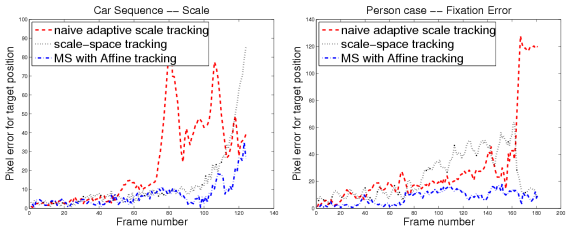
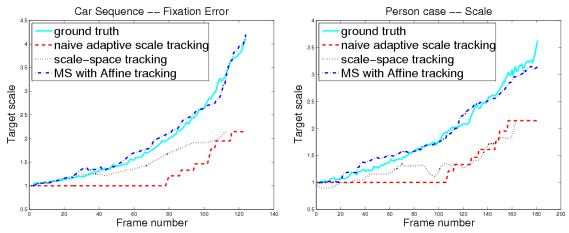


Figure 4. Scale error changes with the feature number reduces

shows how our hybrid model perform on computing scale according to the feature number changing. Those scale errors are obtained by averaging results on several sample sequences. Figure 5 shows fixation error



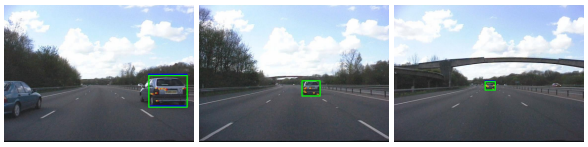
(a) Fixation error between the resulting location and ground truth



(b) The resulting scale relative to the initial frame

Figure 5. Resulting scales and fixation error (measured in pixel) for three trackers. Solid line: ground truth; Dashed line: $\pm 10\%$ adaptive scale MS; Dotted line: scale-space model MS; and dash-dot line: MS with affine hybrid model

for target position and resulting scale for the two sequences above. Figure 5(a) shows the pixel error of the target location from the ground truth labelling by hand, while Figure 5(b) shows the scales computed from the three methods and the ground truth scale, both for the two sequences respectively. These result curves demonstrate that our hybrid model exhibits the lowest location error and the closest scale to the ground truth. The curves representing Collins' scale-space method illustrate that it failed to compute the scale correctly and lost track from frame 114 in left column (for car sequence) and frame 164th in right column (for person sequence). Figure 6 indicates our approach also works well on "re-



(a) car sequence 2: frame #545, #608, #698

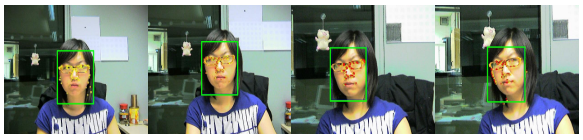
(b) person sequence 2: frame #408, #461, #633

Figure 6. MS with affine hybrid model on target shrink sequences

ceding" targets. The source videos for Figure 6(a) and Figure 6(b) are from PETS2001 (Performance Evaluation of Tracking and Surveillance) database and PETS-CAVIA project data set respectively.

#058, 59ms p=0.27, t=0.53 | #200, 48ms p=-0.98, t=0.67 | #245, 53ms p=-16.25, t=1.8 | #406, 62ms p=14.25, t=9.52

(a) Pan-Tilt control



#054, 47ms p=0, t=0 f=5.76 | #160, 55ms p=4.27, t=1.86 f=6.66 | #250, 71ms p=6.18, t=1.39 f=10.86 | #300, 69ms p=10.45, t=1.75 f=10.2

(b) Pan-Tilt with zoom control

Figure 7. Real-time implements on an active camera. #: frame number; ms: operation time in millisecond; p: pan in degree; t: tilt in degree; f: focal length in millimeter

Our approach is efficient enough for the real-time running. Some resulting frames from two real-time sequences shown in Figure 7, demonstrates the use of the algorithm for closed-loop visual control of pan, tilt and zoom. The camera is working at 15Hz frame rate, and the size of the images are 640x480 pixels. In 7(a), a Pan-Tilt-Unit is controlled to maintain the target in the center of the scene. In 7(b), additionally, zoom control is also applied to preserve the target size.

4. CONCLUSIONS

A new hybrid tracking model is developed for the tracking of non fixed-scale targets. This new method is based on the combination of the Mean-Shift algorithm and the affine Transfer. This hybrid tracker allows for a reliable, view-point invariant and robust tracking. It is fast to follow changes in scale and consistent to predict accurate object window, comparing to other existing methods. We solved the "roam" problem in Mean-Shift by correcting the location according to the candidate window calculated by affinity, even when the scale is underestimated. The hybrid tracking scheme is examined through the comparisons with various existing tracking methods and models. In our implement, the mean cost for the standard MS is 10 millisecond/frame, 100 millisecond/frame for Collins' scale-space method and less than 50 millisecond/frame for our approach. Excellent consistency and agreement with the target motion shown in the comparison experiments indicate that the present hybrid model greatly improves the overall performance in size-varying target tracking.

5. ACKNOWLEDGEMENTS

The authors would like to acknowledge partial support of EC grant IST-027110 for the HERMES project in the EU sixth framework programme.

References

[1] G. R. Bradski. Computer vision face tracking for use in a perceptual user interface. *Intel Technology Journal*, (Q2):15, 1998.

[2] R. T. Collins. Mean-shift blob tracking through scale space. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 234–240, June 2003.

[3] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, volume 25, pages 564–577, May 2003.

[4] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, 24(6):381–395, June 1981.

[5] J. J. Koenderink and A. J. van Doorn. Affine structure from motion. *Journal of the Optical Society of America*, 8(2):377–385, Feb 1991.

[6] T. Lindeberg. Scale-space theory in computer vision, kluwer, dordrecht. *Monograph 1994*, 1994.

[7] P. McLauchlan and J. Malik. Vision for longitudinal vehicle control. In *British Machine Vision Conference*, pages 918–923, September 1997.

[8] I. D. Reid and D. W. Murray. Active tracking of foveated feature clusters using affine structure. *IEEE International Journal of Computer Vision*, 18(1):41–60, April 1996.

[9] E. Rosten and T. Drummond. Fusing points and lines for high performance tracking. In *IEEE International Conference on Computer Vision*, pages II: 1508–1515, 2005.

[10] L. S. Shapiro. *Affine analysis of image sequences*. Cambridge University Press, 1995.

[11] B. Tordoff. *Active control of zoom for computer vision*. PhD thesis, Department of Engineering Science, University of Oxford, Oxford, UK, 2002.