# SEGMENTATION-BASED OBJECT TRACKING USING IMAGE WARPING AND KALMAN FILTERING

Yu Huang*, Thomas S. Huang*, Heinrich Niemann+
* IFP, Beckman Institute, UIUC, Urbana, IL61801
+ Chair for Pattern Recognition, U. of Erlangen-Nuremberg, Germany, 91058
e-mail: {yuhuang, huang}@ifp.uiuc.edu, niemann@informatik.uni-erlangen.de.

## ABSTRACT

*We propose a segmentation-based method of object tracking using image warping and Kalman filtering. The object region is defined to include a group of patches, which are obtained by a watershed algorithm. In a robust M-estimator framework, we estimate dominant motion of the object region. A linear Kalman filter is employed to predict the estimated affine motion parameters based on a second order kinematic model. Image (affine) warping is performed to predict the object region in the next frame. Warping error of each watershed segment (patch) and its rate of overlapping with the predicted region are utilized for classification of watershed segments near the object border. Applications of head and hand tracking using this method demonstrate its performance.*

## 1. INTRODUCTION

Tracking objects in image sequences is an important task for vision-based control, human computer interaction (HCI), content-based video indexing and structure from motion etc. A great variety of visual tracking algorithms have been proposed, they can be classified roughly into two categories [5]. The first is the *feature-based* method. A typical instance in this category estimates the 3D pose of a target object to fit into the image features such as contours given a 3D geometric model of the object. The second is the *region-based* method. Compared to the feature-based methods the region-based methods are more robust, insensitive to small partial occlusions. The region-based methods can be subdivided into two groups: the *view-based* method and the *parametric* method. The view-based method finds the best match of a region in a search area with a reference template. The parametric method assumes a parametric model of changes in the target image and computes optimal fitting of the model to pixel data in a region. Our proposed method belongs to the latter type.

### 1.1 Related Work

Below we discuss some related work in the literature of region-based visual tracking.

Shi and Tomasi [11] put forward the criterion of "good features" by its texturedness and used it in affine feature tracking. It was shown this method is robust to partial occlusion. Parry et. al [8] introduced a region-based tracking method, it mainly updated the template by projecting it around the detected position of the target template and considering the overlap between the template and the segmented image. Tracking results demonstrated its good performance.

Hager and Belhumeur [5] developed a general framework for region tracking which includes models for image changes due to motion, illumination and partial occlusion. They used a cascaded parametric motion model and a small set of basis images to account for shading changes, which will be solved in a robust estimation framework to handle small partial occlusion.

Gleicher [4] introduced *difference decomposition* to solve the registration problem in tracking. This idea decomposes the difference image into a linear combination of the difference templates, now estimation for the difference would be linear combination of the corresponding basis vectors. Sclaroff and Isidoro [10] used this idea to for template registration in his region-based non-rigid motion tracking method. In [10] the non-rigid deformation was represented in terms of eigenvectors of a finite element method. The photometric variation is considered by adding new brightness and contrast terms. They used a modified Delaunay refinement algorithm to construct a consistent triangular mesh for the region of the object.

Nguyen and Worring [7] made their contribution by introducing a contour tracking method incorporating static segmentation by the watershed algorithm. Their method utilized kinds of edge maps from motion, intensity and prediction (contour warping) to update the object contour. It was claimed this method yielded robust results.

### 1.2 Review

In this paper we propose a segmenation-based method of motion estimation which undergoes object tracking. The object region is defined to include a group of patches,

which are obtained by a watershed algorithm. In a robust M-estimator framework, we estimate dominant motion of the object region. A linear Kalman filter is employed to predict the estimated affine motion parameters based on a second order kinematic model. Image (affine) warping is performed to predict the object region. Warping error of each watershed segment (patch) and its rate of overlapping with the predicted region are utilized for classification of watershed segments near the object border. Applications of head and hand tracking using our method demonstrate its performance.

The trend of our work is comparable to [7]: we predict and update the object region, instead [7] dealed with the object contour. The idea of "active blob" [10] also discussed the non-rigid deformation: they used Delaunay triangulation of computer graphics to generate some mesh of the object region, instead our method has employed a powerful segmentation tool — watershed.

This paper is structured as follows. In Section 2, we develop the framework for region-based object tracking. Tracking results are shown in Section 3, and Section 4 will make conclusions.

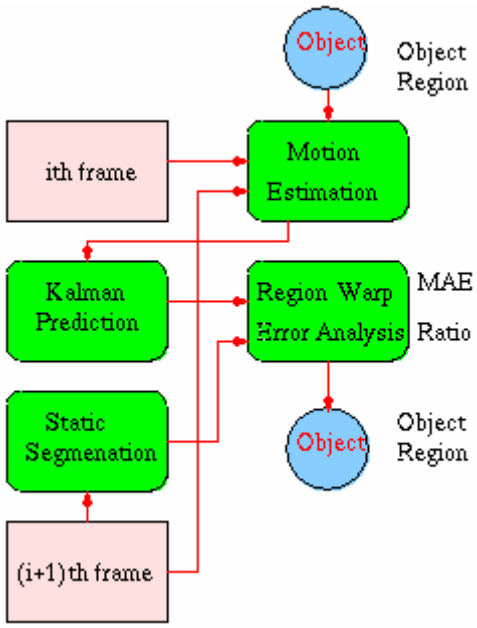



Figure 1 Flow chart of our tracking method

## 2. SEGMENTATION-BASED TRACKING

We assume the object region in the image has been detected in the first frame (detection and location of moving objects in an image sequence is another topic), now the tracking process starts. The flow chart of our method is illustrated in Figure 1. The details of each module are given below.

### 2.1 Motion Estimation using the M-estimator

Here we describe the problem as follows: the inter-frame motion is defined as

$$f(\mathbf{x}, t+1) = f(\mathbf{x} - \mathbf{u}(\mathbf{x};\mathbf{a}), t), \qquad (1)$$

with $f(\mathbf{x}, t)$ as the brightness function in time instant $t$, $\mathbf{x} = (x, y)$ as the coordinate of the image pixel, and $\mathbf{u}(\mathbf{x};\mathbf{a})$ as the motion vector. Without loss of generality, we simply select affine transform as the motion model,

$$\mathbf{u}(\mathbf{x};\mathbf{a}) = \begin{bmatrix} u(x,y) \\ v(x,y) \end{bmatrix} = \begin{bmatrix} a_0 + a_1 x + a_2 y \\ a_3 + a_4 x + a_5 y \end{bmatrix}, \qquad (2)$$

where $\mathbf{a} = (a_0, a_1, a_2, a_3, a_4, a_5)^T$ are the parameters of the affine model. So, the dominant motion estimation of the given region $R$ is formulated as the following robust M-estimator,

$$\min_{(u,v)} E_D = \sum_{(x,y)\in R} r(uf_x + vf_y + f_t, s), \qquad (3)$$

here $f_x$, $f_y$, $f_t$ is partial derivatives of brightness function with respect to $x$, $y$ and $t$, the $r$ - function is chosen as the Geman-McClure function [2] and $s$ is the scale parameter. To solve the problem, there are two different ways to find robustly the motion parameters: one is gradient-based, like the SOR method in [2], another is least squares-based, such as the Iterative Weighted Least Squares (IWLS) method [6]. We test both iteration methods and find the latter one is more stable.

The algorithm begins by constructing the Gaussian pyramid (we set up three levels). When the estimated parameters are interpolated into the next level, they are used to warp (realized by bilinear interpolation) the last frame to the current frame. In the current level only the change are estimated in the iterative update scheme.

### 2.2 Kalman Filtering for Motion Prediction

Kalman filtering is a technique for temporal association and integration in tracking [1]. Based on a second order kinematic model, we can model the affine motion vector evolution as a linear system described by the following equations:

$$\begin{cases} \mathbf{s}_k = A\mathbf{s}_{k-1} + n_{k-1} \\ \mathbf{o}_k = H\mathbf{o}_k + z_k \end{cases}, \qquad (4)$$

with $\mathbf{s}_k$ as the state vector describing the affine motion vector, its first derivative and its second derivative, $v_k$ as the model noise, $\mathbf{o}_k$ as the observation (affine motion) vector and $x_k$ as the observation noise. State matrix $A$ and observation matrix $H$ come from the second order kinematic model. The result in Section 2.1 will input this Kalman filter, which will output a motion prediction result from update of the state vector. This allows the filter to integrate over time the temporal information of the tracked object [1].

### 2.3 Segmentation by Immersion Simulations

In static segmentation, the watershed algorithm of mathematical morphology is a powerful method. Early watershed algorithms are developed to process digital elevation models and are based on local neighborhood operations on square grids. Some approaches use "immersion simulations" to identify watershed segments by flooding the image with water starting at intensity minima [12]. Improved gradient following methods are devised to overcome plateaus and square pixel grids [3]. Here we use the former method.

A severe drawback to the computation of watershed algorithm is over-segmentation. Normally watershed merging is performed along with the watershed generation. But here over-segmentation is welcome, so during tracking we omit the merging process, which saves some computation costs.

### 2.4 Image Warping and Region Anlaysis

Once the predicted motion parameters have been obtained from Kalman filtering, we warp the object region from the *i*th frame to the (*i*+1)th frame. Then the warped region is used to determine which watershed segments enter it according to the following measure: Given that the number of pixels belonging to the warped template in the sub-region (watershed segment) $R_i$ is $Cp_i$ and the number of all pixels in $R_i$ is $C_i$, a ratio $r_i$ is computed,

$$r_i = Cp_i / C_i. \qquad (5)$$

Based on this measure, we discuss further the classification problem of each subregion in these following cases:

1) When $r_i \geq$ r0 (in this paper r0 = 0.9), classify $R_i$ as part of the final object template;

2) When r0 > $r_i \geq$ r1 (here r1 = 0.4), another measure as MAE (Mean Absolute Error) of difference between the warped frame and the current frame is taken into account,

$$M_i = \sum_{\mathbf{x}\in R_i} \left| f(\mathbf{x}, t+1) - f^w(\mathbf{x}, t) \right| / C_i. \qquad (6)$$

where $f^w(\mathbf{x}, t)$ is the warped image of $f(\mathbf{x}, t)$ using the estimated dominant motion parameters; If the warped error $M_i$ of $R_i$ is smaller enough (less than a given threshold, for instance, 10), $R_i$ is still regarded as part of the updated template; Otherwise, we exclude $R_i$ out of the object region.

3) When $r_i$ < r1, $R_i$ will NOT be included in the updated template.

Figure 2 give an illustration to this process: 2(a) and 2(b) are a pair of consecutive frames. For sake of simplicity, we assume the detected object is equivalent to the real object. 2(c) is the warped object, and 2(d) is static segmentation (in blue). In 2(e) the warped template(enclosed in green), watershed segments and the real object are superimposed to clearly illustrate the comparison. The sub-region enclosed in red agrees with the first case, the sub-region enclosed in yellow agrees with the second case, and the subregion enclosed in brown agrees with the third case. The final object template is shown in 2(f).

In our experiments, it is found the warping error analysis is efficient to avoid some misclassification of small regions near the tracked object in the cluttered background.

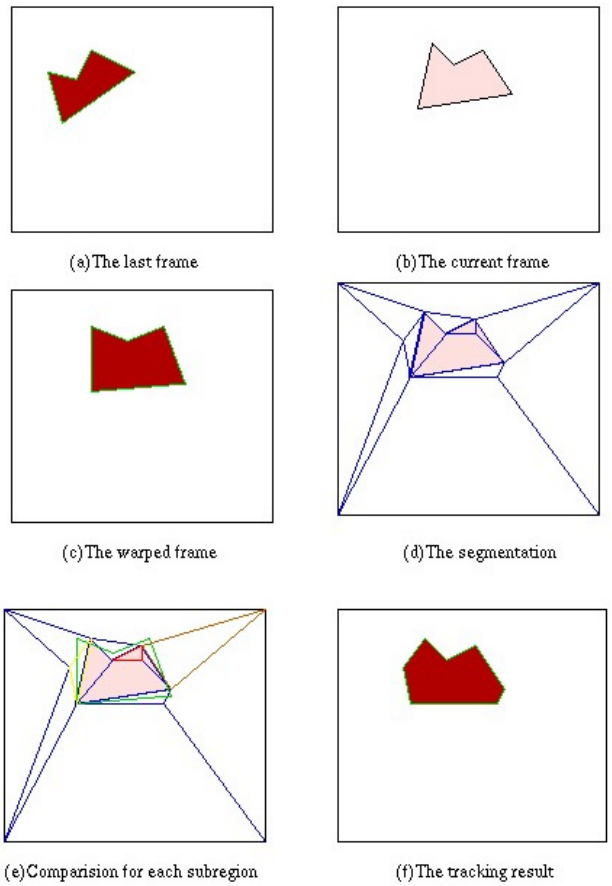


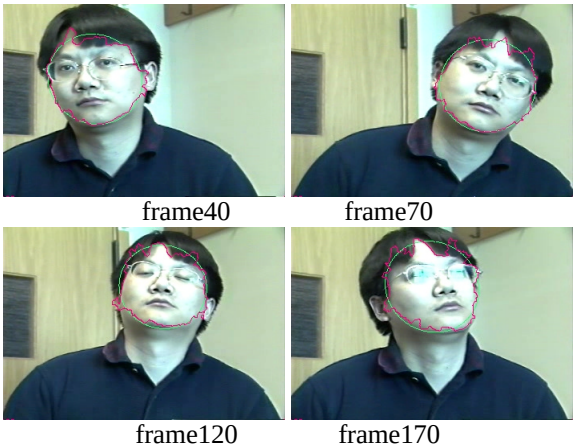

Figure 2 Illustration of template update in tracking

## 3. EXPERIMENTAL RESULTS

In order to illustrate the tracking performance directly, we use an ellipse to approximate the tracked object region. We realize this approach in Visual C++ on Pentium II 400M. Now the processing speed is about 1 seconds per frame if the image size is 350x240. At the initialization, once we manually put an ellipse (in green) on the tracked object, the tracking procedure starts from the marked ellipse region. In our experiments, we focus on head and hand tracking in natural environments.

In Figure 3 a sequence "Head" of 300 frames is used for tracking. In this video sequence the person head is moving

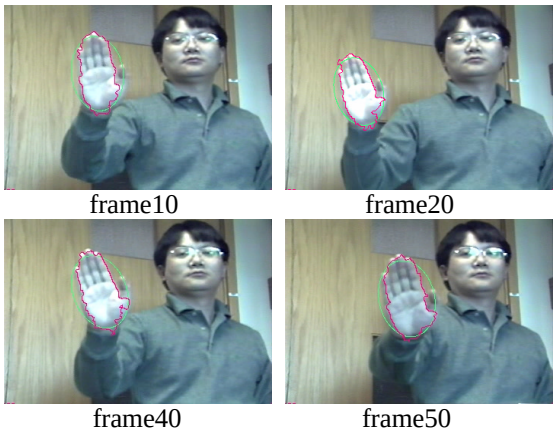
with translation and rotation. We depict the tracked object contours (in red) and its corresponding ellipses (in green) simultaneously on each frame. It is shown the performance of our method in this indoor environment is not bad. Figure 4 shows the hand tracking result from a "Hand" sequence (80 frames), where the hand goes closer to and away from the camera repeatedly. It is obvious we capture the changes of the hand shape.



frame40 frame70

frame120 frame170

Figure 3 Head Tracking Results



frame10 frame20

frame40 frame50

Figure 4 Hand Tracking Results

## 4. CONCLUSION

We put forward a segmenation-based method of motion estimation which undergoes object tracking, its character is full utilization of the object spatio-temporal information in tracking. The image warping based on the predicted affine parameters obtained from linear Kalman filtering only gives a prediction to the object region, and the comparison of each sub-region with the warped template is able to modify the predicted region. Results of head and hand tracking using our method are encouraging.

The disadvantages of our method are also clear in the experiments. First, we rely on the motion estimation of the tracked object; Even though the IWLS method is more stable, we still confronted the divergence in iterations. Second, while we introduce the static segmentation result our method has strong dependence on the performance of the employed watershed algorithm; We expect the images in the sequence are with enough resolution, and the motion blur are also not welcome. Third, due to the object kinematic characters, Kalman filtering does not always provide a correct prediction, in this case we can regard it as a initial guess for motion parameters in the next time instant.

In future, we will consider the variations of illumination [5, 10], which is also an important factor in tracking.

## ACKNOWLEDGEMENT

This research is supported partially by the NSF Grants CDA96-24396, EIA-99-75019 and IIS-00-85980.

## REFERENCE

[1] Bar-Sharlom Y, Fortmann T E, Tracking and Data Association, Academic Press, Inc., 1988.

[2] Black M, Yacoob Y, "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion", ICCV'95, 1995.

[3] Gauch J, „Image segmentation and analysis via multiscale gradient watershed hierarchies", *IEEE T-IP*, 8(1): 69-79, 1999.

[4] Gleicher M, "Projective registration with difference decomposition", IEEE CVPR'97, pp331-337, 1997.

[5] Hager G. and Belhumeur P., "Efficient region tracking with parametric models of geometry and illumination". *IEEE Trans. PAMI*, 20(10):1025-1039, 1998.

[6] Holland P, Welsch R, 'Robust regression using iteratively reweighted least squares', Comm. Statist. Theor. Methods, 1977.

[7] Nguyen H., Worring M., "Multifeature object tracking using a model-free approach", IEEE CVPR, pp 145 –150, 2000.

[8] Parry et. al, "Region Template Correlation for FLIR Target Tracking", British Machine Vision Conference, 1996.

[9] Saber E, Tekalp A, "Face detection and facial feature extraction using color, shape and symmetry-based cost functions", ICPR'96, pp654-658, 1996.

[10] Sclaroff S. and Isidoro J., "Active blobs", ICCV'98, 1998.

[11] Shi J. and Tomasi C, "Good features to track". In *Proc. Computer Vision and Pattern Recognition*, 1994.

[12] Vincent L, Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations", *IEEE T-PAMI*, 13(6): 583-589, 1991.