Robust Foveal Wavelet-Based Object Tracking

Aldo Maalouf, *Member IEEE*, Mohamed-Chaker Larabi, *Senior Member IEEE*
XLIM-SIC Laboratory, UMR CNRS 6172, University of Poitiers
{maalouf, larabi}@sic.univ-poitiers.fr

***Abstract*—In this work, a foveal wavelet-based Mean Shift Tracking Algorithm is presented. The foveal wavelets introduced by Mallat [16] are known by their high capability to precisely characterize the holder regularity of singularities. Therefore, by using the foveal wavelet transform, image features are accurately identified and are well discriminated from noise. These wavelets are used to extract the texture features of the target object. The extracted features are then used to construct a joint color-foveal textures histogram to represent the target object. Once the joint histogram is obtained, it is applied to the mean shift framework in order to track a target object in a video sequence. The experimental results showed that the proposed approach overcomes the traditional mean shift tracking technique as well as other existing tracking algorithms.**

***Keywords*-Object Tracking, Foveal Wavelet, Mean Shift Algorithm.**

I. Introduction

The problem of tracking video objects is one of the major issues in video surveillance systems. The challenges in tracking include occlusion, illumination change, object perspective or scale change, etc. Among the existing tracking algorithms, the mean shift algorithm, that was firstly proposed by Fukunaga et al. [7], has become popular due to its efficiency and simplicity. Beside the mean shift algorithm there have been too many tracking approaches, or even modified mean shift algorithms, in the literature that addressed the problems of occlusion and illumination or scale changes in object tracking. Let us first take a look to the most commonly used tracking algorithms.

In [15], Lucas et al. proposed a KLT-based tracking method that establishes correspondence by using a gradient-based minimisation technique. In their approach, Lucas et al. considered that the tracked objects are image patches with two high eigenvalues of the structure tensor as depicted in [18]. These eigenvalues are the same as regions centered around Harris interest points [9]. These Harris regions possess the property that they are different from all regions in their neighbourhood, a necessary condition for establishing reliable point-to-point correspondence. However, these points do not have any property that would facilitate the minimisation step.

Other existing tracking approaches, for instance the approach of Castle et al. and the one proposed by Klein et al. [2] [14], are based on fast detectors and establish correspondence by matching of detected points, which is feasible if only a small number of alternatives must be verified, i.e. when the error in prediction and hence the search area is small with respect to the density of points.

This work was supported by the French National Research Agency (ANR) under the QuIAVU project

To design more robust object tracking algorithms, some authors combined MS (mean shift) with local point feature-based tracking or Bayesian framework like PFs (particle filters). In [12], Zhou et al. proposed an expectation-maximization algorithm that integrates SIFT features along with color-based object appearance in MS. In [1], Chen et al. used feature points to handle occlusion and scaling under the MS framework.

In [20], Khan et al. used the MS framework along with consensus point feature correspondences in order to improve tracking accuracy via a coarse-to-fine process. In [5], Shan et al. proposed an approach in which the MS is embedded in particle filters to track human hands. In [19], Zhong et al. proposed to weight particles by using an observation model. Thereafter, they applied the MS on particles with large weights, called elite particles. In [13], Khan et al. combines PFs and anisotropic MS seeking multiple appearance modes by partitioning a rectangular bounding box into sub-regions.

However, the above appearance based tracking methods share a common drawback which is their inefficiency to perform tracking in noisy video sequences and when the target is having similar appearance to the background. For this latter, many authors proposed to combine the color histogram with edge features [4] [8] [17].

Nevertheless, to our knowledge, none of these methods addressed the problem of object tracking in noisy sequences because intensity-based edge and feature detectors cannot distinguish between various transition types. For that, our attention is directed toward the use of multiscale approaches, namely the wavelets. The choice of which wavelet to use should be based on the precise contour localisation capability of that wavelet.

The foveal wavelets were first introduced in [16]. They mimic the non uniform distribution of photoreceptors on the retina. The visual resolution is highest at the center (fovea) of the retina, but falls off away from the fovea. This effect is modeled by foveal approximation spaces introduced in [16]. Projections in a foveal approximation space approximate functions with a resolution that decreases proportionally to the distance from the point of interest. Foveal wavelet coefficients give a pointwise characterization of edges.

Having well represented the edges, a joint color-texture histogram is constructed. This joint histogram substitutes the traditional color-histogram that is used to represent the target in the traditional MS approach. The proposed foveal wavelet-based target representation allows a better representation of the target structural information even in noisy sequences. Consequently, a better target tracking is achieved.

The paper is organised as follows. In section 2, a review of the foveal wavelet is given. Section 3 aims to represent the

traditional MS approach. In section 4, our tracking approach is represented. In section 5 experimental results are shown. Finally, section 6 concludes the paper.

II. Foveal Wavelets

A. Review of the foveal wavelets

Edges are considered as one-dimensional singularities that move in the plane of the image. Foveal wavelets are orthogonal wavelets that are centered at the same location as if to absorb the singular behavior of the frame image [16]. These wavelets zoom on a single position u. If $\mathbf{V}_u$ is the space generated by the foveal wavelet located at u, then the orthogonal projection of a function f in $\mathbf{V}_u$ is given by:

$$P_{V_u} f(t) = \sum_{j=-\infty}^{J} \sum_{k=1}^{2} \left\langle f, \psi_{j,u}^{k} \right\rangle \psi_{j,u}^{k}(t) \tag{1}$$

Where Ψ is the mother foveal wavelet. These wavelets are characterized by their ability to eliminate singularities located at u. If f is differentiable in a left and right neighborhood of u, but not in u then it has been shown in [16] that $fP_{V_u}f$ is continuous at u and has a bounded derivative over a whole neighborhood of u. Therefore, the singularity of f at u is absorbed by the wavelet coefficients at u. Singularities of f are entirely characterized by the foveal wavelet coefficient at u. Singularities can be detected by computing

$$\varepsilon(u) = \sum_{j=-\infty}^{J} 2^{-2j} \sum_{k=1}^{2} \left| \left\langle f, \psi_{j,u}^{k} \right\rangle \right|^2 \tag{2}$$

If f has a Lipschitz regularity $\alpha < 1$ at u, and hence is not differentiable at u then $\varepsilon(u) \to +\infty$, but if f has Lipschitz regularity $\alpha > 1$ at u then $\varepsilon(u) < +\infty$. Singularities are thus detected from the amplitude of $\varepsilon(u)$.

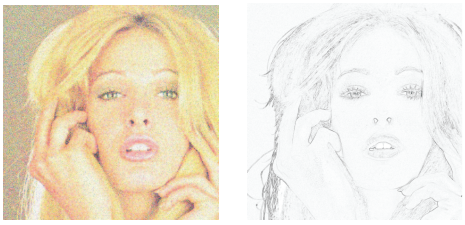
We can, therefore, distinguish between noise singularities (negative Lipschitz component) and edge singularities from the amplitude of $\varepsilon(u)$.

B. Edge detection using foveal wavelets

The singularities of a frame I are detected with one dimensional foveal wavelets, along each line and each column of the frame image. Detected singularities are chained together to form edge curves in two dimension. Let $\{f(x_1, u_2)\}_{x_1 \in \mathbb{R}}$ be a horizontal scan line, where u_2 is fixed and x_1 varies. $\{f(x_1, u_2)\}_{x_1 \in \mathbb{R}}$ is decomposed over one-dimensional foveal wavelets. For each u we compute

$$\varepsilon_{u_2}(u) = \sum_{j} 2^{-2j} \sum_{k=1}^{2} \left| \left\langle f(x_1, u_2), \psi_{j,u}^{k}(x_1) \right\rangle \right|^2 \tag{3}$$

Any singularity corresponds to a point u_1 where $\varepsilon_{u_2}(u)$ is locally maximum when u varies. This singularity is located in the frame image plane at the position (u_1, u_2). The same procedure is repeated along the columns of the image to detect singularities. Horizontal and vertical detected singularities are chained together to form edge curves. Figure 1 shows a family of curves detected along each row and column of a noisy image. This representation is used to construct the joint color-edge histogram in the MS framework.



(a) (b)

Fig. 1. (a) noisy image, (b) edge curves detected along each row and column

III. Mean Shift Algorithm

The mean-shift algorithm is a non-parametric density gradient estimator. It is basically an iterative expectation maximization clustering algorithm executed within local search regions. Comaniciu has adapted the mean-shift for the tracking of manually initialized targets [4]. The mean-shift tracker provides accurate localization and it is computationally feasible. A widely used form of target representation is color histograms, because of its independence from scaling and rotation and its robustness to partial occlusions. Define the target model as its normalized color histogram $q = \{q_r\}_{1,\ldots,m}$

$$q_r = C \sum_{i=1}^{n} k\left(\|x_i^*\|^2\right) \delta\left[b(x_i^*) - r\right] \tag{4}$$

where m is the number of bins. The normalized color distribution of a target candidate $p(y) = \{p_r(y)\}_{1\ldots r_h}$ centered in y can be calculated as

$$p_r(y) = C_h \sum_{i=1}^{n_h} k\left(\left\|\frac{y - x_i}{h}\right\|^2\right) \delta\left[b(x_i) - r\right] \tag{5}$$

where $\{x_i\}_{i=1\ldots n_h}$ are the n_h pixel locations of the target candidate in the target area, $b(x_i)$ associates the pixel x_i to the histogram bin, $k(x)$ is the kernel profile with bandwidth h, and C_h is a normalization function defined as

$$C_h = \frac{1}{\sum_{i=1}^{n_h} k\left(\left\|\frac{y - x_i}{h}\right\|^2\right)} \tag{6}$$

In order to calculate the likelihood of a candidate, we need a similarity function which defines a distance between the model and the candidate. A metric can be based on the Bhattacharyya coefficient, defined between two normalized histograms $p(y)$ and q as

$$\rho\left[p(y), q\right] = \sum_{r=1}^{m} \sqrt{p_r(y)\, q_r} \tag{7}$$

Hence the distance is defined by

$$d\left[p(y), q\right] = \sqrt{1 - \rho\left[p(y), q\right]} \tag{8}$$

To track the target using the Mean Shift algorithm, we iterate the following steps:

1) Choose a search window size and the initial location of the search window.
2) Compute the mean location y_0 in the search window.

3) Center the search window at the mean location computed in Step 2 and compute the Bhattachayya coefficient $P_r(y_0)$ using (7) :

$$\rho\left[p\left(y\right),q\right] \approx \frac{1}{2}\sum_{r=1}^{m}\sqrt{p_r\left(y_0\right)q_r} + \frac{1}{2}C_h\sum_{i=1}^{n_h} w_i k\left(\left\|\frac{y-x_i}{h}\right\|^2\right) \tag{9}$$

where

$$w_i = \sum_{r=1}^{m}\sqrt{\frac{q_r}{p_r\left(y_0\right)}}\delta\left[b\left(x_i-r\right)\right] \tag{10}$$

4) Repeat Steps 2 and 3 until convergence (or until the mean location moves less than a preset threshold). The estimated target moves from y to a new position y_1 defined by:

$$y_1 = \frac{\sum_{i=1}^{n_h} x_i w_i}{\sum_{i=1}^{n_h} w_i} \tag{11}$$

IV. Foveal Wavelet-Based Mean Shift Algorithm

We made use of the RGB channels and the information provided by the foveal wavelet-based edge representation to jointly represent the target. First we construct the edge map of the target region in a frame I by using the foveal wavelet -based procedure presented in section II. Then, for each edge pixel in the target region, we compute the Lipschitz component α using the procedure explained in [16]. Thereafter, we compute the orientation at each edge point by using the following equation:

$$\theta\left(x,y\right) = \tan^{-1}\left(\frac{I\left(x,y+1\right)-I\left(x,y-1\right)}{I\left(x+1,y\right)-I\left(x-1,y\right)}\right) \tag{12}$$

To obtain the color and texture distribution of the target region, we use (5) to compute the distribution for the target model q, in which $r = 8\times 8\times 8\times 3\times 3$. The first three dimensions represent the quantized bins of the RGB channels, the forth dimension represent the quantized bins of the orientation, while the last dimension is for the quantized bins of the Lipschitz component. We have used 8 orientations included in $\left[-\frac{\pi}{2},\frac{\pi}{2}\right]$ and 8 Lipschitz components included in $[-1.5,1.5]$.

Our tracking algorithm is summarized as follows:

It is to be noted that the use of the foveal wavelets based edge detection permits to perform tracking in noisy sequences. Furthermore, the use of the Lipschitz component permits to distinguish between strong and weak edges and consequently, to achieve better tracking when the background and the target object have similar appearance.

V. Experimental Results

We perform experiments to validate our tracking algorithm and compare it with existing ones. To achieve this, we compare our approach with the ones proposed by Comaniciu et al. [3] and Ning et al. [17]. In our experiments, we have pre-selected image region to be tracked in the following video sequences.

INPUT: a threshold ϵ, the target model q and its central location y_0 in the previous frame.

Step 1: In the current frame compute the foveal wavelet transform of the target region.
Step 2: Detect the edges in the target region using (3).
Step 3: For each edge pixel compute the Lipschitz component as explained in [16].
Step 4: For each edge pixel compute the orientation using (12).
Step 5: Compute the distribution of the target candidate $P_r(y_0)$ using (5).
Step 6: Compute the weights w_i using (10).
Step 7: Compute the new central location using (11).
Step 8: Let $D = \|y_1 - y_0\|$,
if $D < \epsilon$
Goto step 9
else
Goto step 1
Step 9: Load the next frame with an initial location $y_0 = y_1$.
Step 10: **Goto** step 1

The first sequence *tropical* is taken from the database [10], shows a soldier walking in a forest. The region of interest (the soldier) has similar colour to that of the background and many objects occlude temporarily the tracked region. The second sequence, *man* [11], shows a person walking along a car park. Apart from object's colour similarity to the nearby cars and the shadowed areas, the video contains numerous instabilities, resulting from a shaking camera, fast zoom-ins and zoom-outs, and a wide range of a view angles. The last sequence, *otcbvs* is a part of a multimedia benchmark dataset collection [6] to which we have added a White Gaussian Noise with a variation of 0.02. In this sequence, the small-sized region of interest undergoes significant intensity changes as it enters the shadowed areas of the walkway and the entrance of a building. In order to objectively evaluate the performance of the developed tracking technique, we made use of the Root Mean Square Error (RMSE) performance measure:

$$RMSE(z) = \sqrt{\left(x_z-\tilde{x}_z\right)^2+\left(y_z-\tilde{y}_z\right)^2} \tag{13}$$

where $(\tilde{x}_z,\tilde{y}_z)$ is the upper left corner coordinates of the tracking box and (x_z,y_z) is the corresponding ground truth generated manually. The mean of RMSE is presented in Table I.

As we can see from the performance measures shown in Table I, our foveal wavelet based approach outperforms the other two approaches in term of accuracy. It also appears to be more stable than the other two methods in the presence of noise. This is due to the fact that foveal wavelets were capable to characterize the object features in spite of the presence of noise.

It can be also observed from the 'difficult' frames shown in Figure 2 that our algorithm performs better than the other two techniques. In fact these frames offer more insight into the performance and robustness of the tested techniques. In the *tropical* sequence, the MS algorithm [3] failed to track the

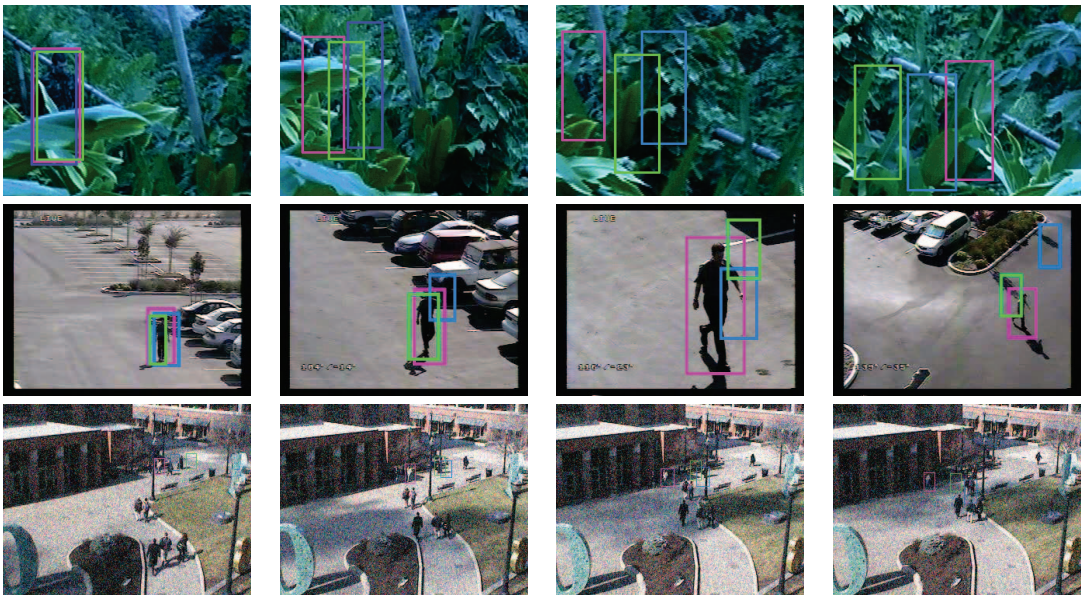


Fig. 2. Video frames with tracker output (our method in Magenta, the method of Comaniciu et al. [3] in blue and the method of Ning et al. [17] in green)

Seq. Name	our method	Ning et al. [17]	Comaniciu et al. [3]
Tropical	4,21	17,12	32,12
man	6,22	10,02	19,22
otcbvs	8,2	37,74	45,32

TABLE I
MEAN RMSE (PIXELS)

soldier in the frames where the object and the background have approximatively the same appearance. Our algorithm achieved a better accuracy in tracking the soldier than the one proposed in [17]. In *man* sequence, the three tested algorithms manage to track the selected target. However, our method identifies the scale and the position of the object with the best accuracy. Unlike the other two methods, the one proposed in this paper tracks the object throughout the noisy *otcbvs* sequence demonstrating the robustness to noise.

VI. CONCLUSION

In this paper, a foveal wavelet based mean-shift tracking algorithm is proposed. The proposed method makes use of foveal wavelet to accurately characterize objects features and perform more accurate tracking. Experimental results showed that the proposed method achieves better accuracy and robustness to noise while tracking object than the original MS algorithm and other existing methods. The developped approach will be included in a videosurveillance monitoring system allowing to have a quality-oriented tecking of objects.

REFERENCES

[1] A.Chen, M.Zhu, Y.Wang, and C.Xue, *Mean shift tracking combining sift*, in Proc. ICALIP, 2008.

[2] R. O. Castle, G. Klein, and D. W. Murray, *Video-rate localization in multiple maps for wearable augmented reality*, in 12th IEEE Int Symp on Wearable Computers, 2008, pp. 15–22.

[3] D. Comaniciu and P. Meer, *Mean shift: A robust approach toward feature space analysis*, IEEE Trans. Patt. Anal. Mach. Intell **22** (2002), no. 5, 603–619.

[4] D. Comaniciu, V. Ramesh, and P. Meer, *Kernel-based object tracking*, IEEE Trans. Patt. Anal. Mach. Intell **25** (2003), no. 5, 564–575.

[5] C.Shan, Y.Wei, T.Tan, and F.Ojardias, *Real time hand tracking by combining particle filtering and mean shift*, Proc. ICAFGR, 2004.

[6] J. Davis and V. Sharma, *Otcbvs benchmark dataset collection*.

[7] K. Fukunaga and L. D. Hostetler, *The estimation of the gradient of a density function, with applications in pattern recognition*, IEEE Trans. Information Theory **21** (1975), no. 1, 32–40.

[8] I. Haritaoglu and M. Flickner, *Detection and tracking of shopping groups in stores*, Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2001.

[9] C. Harris and M. Stephens, *A combined corner and edge detector*, in Proc. of 4th Alvey Vision Conference, 1988, pp. 147–151.

[10] http://www.imagefusion.org, *The eden project multi-sensor data set*, (2006).

[11] http://www.perceptivu.com, *Target tracking movie demos*.

[12] H.Zhou, Y.Yuan, and C.Shi, *Kernel-based method for tracking objects with rotation and translation*, in IJCV'08, 2008.

[13] Z. Khan, I. Gu, and A.Backhouse, *Joint particle filters and multimode anisotropic mean shift for robust tracking of video objects with partitioned areas*, Proc. ICIP, 2009.

[14] G. Klein and D. Murray, *Parallel tracking and mapping for small ar workspaces*, in ISMAR'07, 2007.

[15] B. D. Lucas and T. Kanade, *An iterative image registration technique with an application to stereo vision*, in IJCAI'81, 1981, pp. 674–679.

[16] S. Mallat, *Foveal orthonormal wavelets for singularities*, Tech. report, Ecole Polytechnique, 2000.

[17] J. Ning, L. Zhang, D. Zhang, and C. Wu, *Robust object tracking using joint color-texture histogram.*, IJPRAI.

[18] J. Shi and C. Tomasi, *Good features to track*, in CVPR'94, 1994, pp. 593–600.

[19] S. Zhong and F. Hao, *Hand tracking by particle filtering with elite particles mean shift*, Proc. IWFCST, 2008.

[20] Z.Khan, I. Gu, T.Wang, and A.Backhouse, *Joint anisotropic mean shift and consensus point feature correspondences for object tracking in video*, in Proc. ICME, 2009.