World Applied Sciences Journal 10 (Special Issue of Computer & Electrical Engineering): 20-31, 2010
ISSN 1818-4952
© IDOSI Publications, 2010

Robust Object Tracking in Crowded Scenes Based on the Undecimated Wavelet Features and Particle Filter

[1]*Hamed Moradipour,* [1]*Hossein Ashtiani and* [2]*Amir Aliabadian*

[1]Sama Technical and Vocational Traning School, Islamic Azad University, Babol Branch, Babol, Iran
[2]Faculty member of Electronic & Electrical Department, Shomal University

Abstract: A Scale Invariant Feature Transform (SIFT) based on particle filter algorithm is presented for object tracking. We propose a new algorithm for object tracking in crowded video scenes by exploiting the properties of Undecimated Wavelet Packet Transform (UWPT) and particle filter. SIFT features are used to correspond the region of interests across frames. Meanwhile, feature vectors generated via the coefficients of the UWPT is applied to conduct similarity search that is based on particle filter. The advantage of using structural similarity index UWPT domain is that it allows spatial translations, rotations and scaling changes. Experimental results show that the proposed algorithm has good performance for object tracking in noisy crowded scenes on stairs, in airports, or at train stations in the presence of object translation, rotation, small scaling and occlusion.

Key words: Object tracking . sift feature . undecimated wavelet transform . particle filter

INTRODUCTION

As an active research topic in computer vision, visual surveillance in dynamic scenes attempts to detect, recognize and track certain objects from image sequences. Tracked objects in video sequences can be used for many applications such as video surveillance, visual navigation and monitoring, content-based indexing and retrieval, object-based coding, traffic monitoring, sports analysis for enhanced TV broad-casting and video postproduction. The goal of object tracking is to determine the position of the object in images continuously and reliably against dynamic scenes [1]. To achieve this target, a number of elegant algorithms have been established.

Two major components can be distinguished in a typical visual tracker. *Target Representation and Localization* is mostly a bottom-up process which has also to cope with the changes in the appearance of the target. *Filtering and Data Association* is mostly a top-down process dealing with the dynamics of the tracked object, learning of scene priors and evaluation of different hypotheses. For example considering Gaussian and linear problems, Welch and Bishop [2] presented a Kalman filter-based method for tracking a user's pose for interactive computer graphical. The proposed single-constraint-at-a-time (SCAAT) tracking utilized single observations from optical sensors and fused the measurements from different sensors in order to improve the tracking accuracy and Stability. As a promising solution to non-Gaussian and non-linear systems, particle filter-based approaches have been included in current tracking technologies. These schemes [3, 4] recruited particles for computing a sampled representation of the posterior probability distribution over scene properties of interest, based on image observations. Other tracking strategies can also be found as Multiple Hypothesis Tracking [5, 6], kernel-based tracking [7, 8] and optical flow-based tracking [9].

Both target localization and registration maximizes a likelihood type function. In mean shift tracking algorithms, a color histogram is used to describe the target region. The Kullback-Leibler divergence, Bhattacharyya coefficient and other information-theoretic similarity measures are commonly employed to measure the similarity between the template (or model) region and the current target region. Tracking is accomplished by iteratively finding the local minimum of the distance measure functions.

The method in [10] uses a particle filter to track global statistics of object shape and color. Inserting color to the state of the particle filter yields robustness to background clutter and occlusions. Sampling in the state space, however, is rather expensive. In fact, the clutter problem can also be overcome by using a more discriminative appearance model.

A HMM is normally used to extract the transformation between two images or moving 3D structures in object tracking. However, this is not deterministic and since the model is hidden, there may in fact be more than one possibility of transformation that results in the feature positions. Thus the most likely sequence of transitions is sought. Using algorithms such

Corresponding Author: Mr. Hamed Moradipour, Sama Technical and Vocational Traning School, Islamic Azad University, Babol Branch, Babol, Iran

as the Baum-Welch algorithm and its modifications [11], people train HMM by adjusting the weights of the transitions to better model the relationship of the actual training samples. HMM-based approaches do not require analytical solutions to certain problems, being effective in handling very complicated environments. Nevertheless, the required training stage in HMM must be supervised and it is difficult to apply a pre-trained HMM for the overall applications. Similarly, an ANN also needs to determine its weights by training, although ANN methodology has been optimistically applied to object (or motion) tracking [12, 13].

Most of the existing algorithms are unable to track objects in the presence of variations in illumination, appearance and camera angle, as most of these algorithms working in spatial domain use features which are sensitive to these variations. In recent years the wavelet feature based techniques have gained popularity in object tracking. One of the features of Discrete Wavelet Transform (DWT) is that the spatial information is retained even after decomposition of an image into four different frequency coefficients. However, one of the major problems with real wavelet transform is that it suffers from shift-sensitivity [14]. Undecimated Wavelet Packet Transform (UWPT) has been used to overcome the problem of shift sensitivity.

The main contribution of the paper is to introduce a new framework for efficient tracking of non-rigid objects. In this work for target representation and localization use the SIFT algorithm. This approach transforms an image into a large collection of local feature vectors, each of which is invariant to image translation, scaling and rotation and partially invariant to illumination changes and affine or 3D projection. For tracking we use the particle filter that it is based on feature vectors generated via the coefficients of the Undecimated Wavelet Packet Transform (UWPT). The key advantage of UWPT is that it is redundant and shift-invariant and it gives a denser approximation to continuous wavelet transform than that provided by the orthonormal discrete wavelet transform [15, 16]. In contrast to the conventional methods for solving the tracking problem that use spatial domain features, it introduces a new transform domain feature-based tracking algorithm that can handle object movements, limited zooming e.ects and, to a good extent, occlusion. Moreover, we have shown that the feature vectors are robust to various types of noise [17, 18].

This paper presents a new algorithm that satisfies the two qualities: simplicity and robustness. Simplicity implies that the algorithm is easy to implement and has the minimum number of parameters. Robustness implies the ability of the algorithm to track objects under difficult conditions which include:

- Occlusions and lighting changes.
- Changing of object orientation or view point.
- Track both rigid and nonrigid objects without any preassumption, training, or object shape model.
- Efficiently track the objects in the crowded video sequences such as crowds on stairs, in airports, or at train stations.
- Robust to different types of noise processes such as additive Gaussian noise
- Partial occlusion of the object can be successfully handled.
- Efficiency tracks the object in a moving camera.

OVERVIEW OF THE UWPT

The discrete wavelet transform gives good frequency selectivity at lower frequencies and good time selectivity at higher frequencies. This tradeoff in the Time-frequency (TF) plane is well suited to the representation of many natural signals and images that exhibit short duration high-frequency and long-duration low-frequency events.

The discrete wavelet transform for a one dimensional signal decomposes the signal into two sub-bands called approximations (resulted from convolving the original signal by a low-pass filter) and details (resulted from convolving the original signal by a high-pass filter). It also decimates the output of filters and hence, at each level of decomposition the length of the signal is halved. Moreover, continuing the decomposition, it decomposes only the approximation sub-band. In order to have a more complete interpretation of the signal behavior a wavelet packet transform is applied to decompose also details, but it still decimates the filters outputs. An undecimated wavelet packet transform repeats the filtering on both the low-pass and the high-pass bands without any down sampling. Therefore, the UWPT expansion is redundant and provides a denser approximation [19, 20]. For a 2D signal (image), the UWPT decomposes each sub band into four sub bands at each level.

The desired transform for object tracking application should be linear and shift-invariant. The wavelet transform, which is both linear and shift-invariant, is the Undecimated Wavelet Packet Transform (UWPT) [21, 22]. Moreover, the UWPT expansion is redundant and provides a denser approximation compared to the approximation provided by the orthonormal discrete wavelet transform. From the implementation point of view in the context of filter banks, in addition to the low pass band, we repeat the filtering on the high pass band without any down sampling (decimation). The result is a complete undecimated wavelet packet transforms. A tree

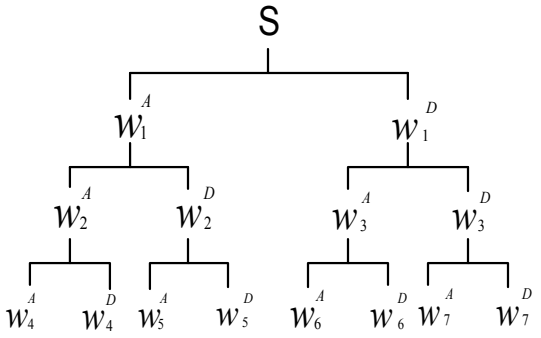


Fig. 1: Undecimated wavelet packet transform tree for one dimensional signal S

representation and sample bands of UWPT are depicted in Fig. 1.

THE PROPOSED ALGORITHM

Overview of the proposed algorithm: The aim of temporal tracking is to locate the object of interest in the successive frames based on the information about the object at the reference and current frames. In our algorithm, object tracking is performed by temporal tracking of a rectangle around the object at a reference frame. A general block diagram of the algorithm is shown in Fig. 2. Initially, the user specifies a rectangle around the boundary of the object at the reference frame. SIFT features are used to correspond the region of interests across frames. This feature is used to localize the search window in the next frame and guide the search window through the tracking process. Then, we have exploited the feature selection in wavelet domain for exactly localized the desired objects in a video sequence. We generate a feature vector and use block matching algorithm in the UWPT domain for exactly localized objects in crowded scenes in presence of occlusion and noise.

The Undecimated Wavelet Transform (UWT) has been used for finding Wavelet Packet (WP) tree. Due to time invariant property of Undecimated Wavelet Packet Transform (UWPT) an object reposition in image will have little impact on the value of wavelet domain coefficients. Also it will ease feature selection procedure, because the sub bands in the decomposition tree will have the same size as input image. In general, Biorthogonal wavelet bases which are particularly useful for object detection [23] could be used for the generation of UWPT tree. We can construct a feature vector that corresponds to each pixel in the region around the object. These FVs can be used to find the best matched region in successive frames; that is, pixels within region *r* are used to find the correct location of the object in frame *t*+1. The process of matching region *r* in frame *t* to the corresponding region in frame *t* c 1 is performed through the full search of the region in a search window in frame *t*+1, which is adaptively determined by the block matching algorithm and Euclidean distances to find the best matched regions.

In the next step, we use particle filter based texture feature cues for tracking. In the particle filter, the weight of each particle is determined by Bhattacharyya coefficient of two corresponding UWPT features. Applying the proposed algorithm results in improved motion tracking which recovers from partial occlusion, rotation and scale.

The entire algorithmic flowchart can be summarized as follows:

- Define a rectangle on the region of interest in the first frame of a video sequence.
- Extracting SIFT features within this region to localize the search window in the next frame.
- Generate an FV for pixels in both region *r* and search window.
- Find the best match for *r* in the search window by calculating the minimum sum of the Euclidean distances between the FVs of the pixels of search regions and FVs of the pixels within region *r*.
- Use particle filter based texture feature cues for tracking.

The procedure to search for the best matched region is the general block-matching algorithm.

In the implementation, the extracting features of the object to be tracked are continuously evaluated. Computational instability may be raised due to lost UWPT or SIFT features (e.g., occlusions). In this case, the estimated probability distribution in the previous frame can be used to dominate locating the object till the object appears again.

The feature vector generation: In the first step, the wavelet packet tree for the desired object in the reference frame is generated by the UWPT. As mentioned in the previous section, the UWPT has two properties that make it suitable for generating invariant and robust features in image processing applications [24, 25].

It has the shift-invariant property. Consequently, feature vectors that are based on the wavelet coefficients in frame *t* can be found again in frame *t*+1, even in the presence of partial occlusion.

All the sub bands in the decomposition tree have the same size equal to that of the input frame (no down sampling), which simplifies the feature extraction process (Fig. 3).

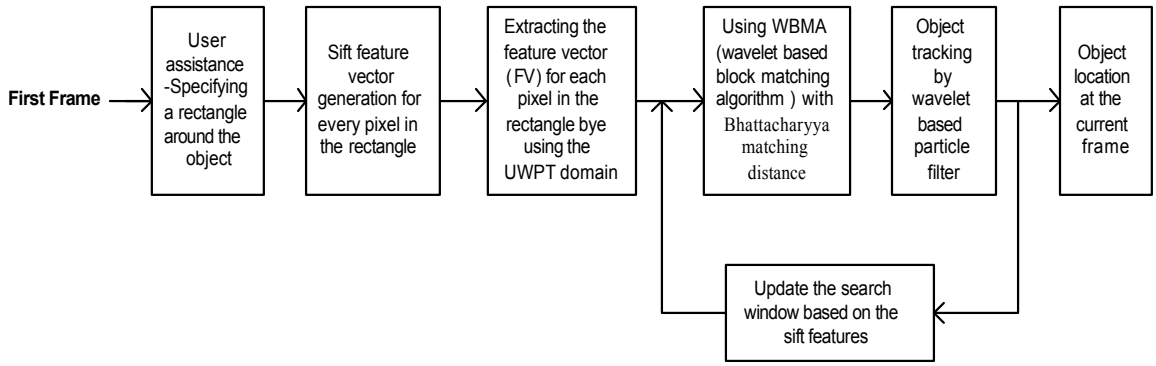


Fig. 2: Block diagram of our proposed algorithm

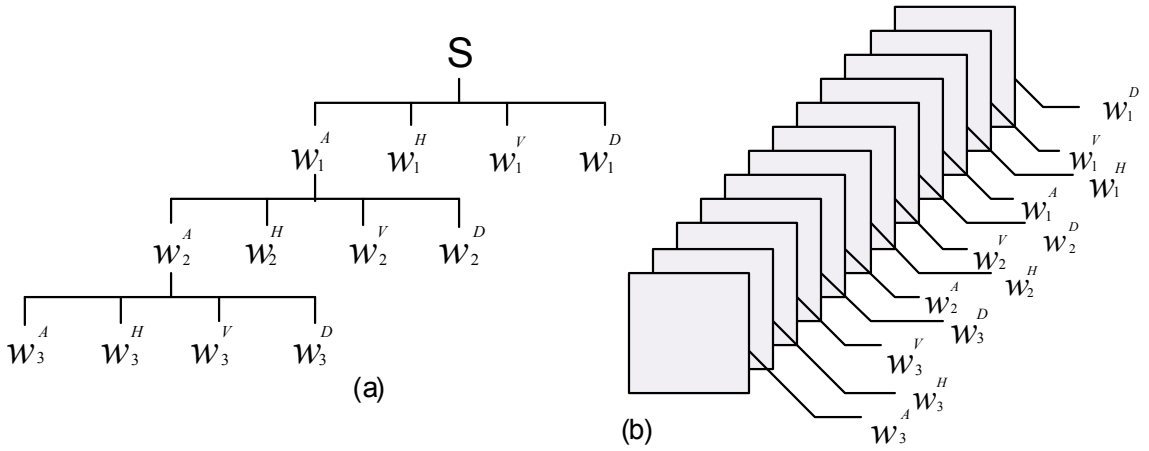


Fig. 3: Feature vector selection: (a) a selected basis tree, (b) ordering of the sub band coefficients to extract the feature vector

Moreover, UWPT alleviates the problem of sub band aliasing associated with the decimated transforms such as DWT. The output of this step is an array of node index numbers of the UWPT tree.

Scale invariant feature transform: The Scale Invariant Feature Transform (SIFT) has become a popular feature extractor for vision-based applications. It has been successfully applied to metric localization and mapping using stereo vision.

SIFT [26] is a method of describing the features of an object such that the same object can be recognized within variance to scale, rotation and affine transformations. The method uses Difference of Gaussian (DoG) to locate points on an object that are stable in scale space and then describe these feature points by the relative gradient orientation of the feature point compared with surrounding points within some window size. This descript tor is made of 128 elements for each feature point using four bins in the x and y directions and eight bins for the orientation. In various applications of SIFT objects are Identified by comparing the number of points that fall within a Euclidean distance threshold between two images. This is an all-to all comparison and has no restrictions on the relative position of points. In this way, if an object is severely occluded, it can still be found in an image if enough of the available feature points are considered a match. This approach works well for computer vision where object recognition can have more broad applications. The SIFT feature descriptor is represented as f = {p, s, o, hist} where p is the 2-D position of the feature in terms of the image coordinate, s is the feature scale, o is the feature vector direction and hist is the gradient orientation distribution quantized into 128 bins.

SIFT-based object recognition is performed by matching each key point extracted from the current image independently to SIFT model built offline.

Feature point selection: The first step in selecting stable points is to find the DoG images. The base image is nominally smoothed using a Gaussian function, Eq. (1), with $\sigma_n = 0.5$, resulting in I(x,y).

$$G(x,y)=\frac{1}{2\pi\sigma^2}e^{-(x^2+y^2)/2\sigma^2} \quad (1)$$

The first Gaussian image is created using $g_\sigma=\sqrt{\sigma_0^2+\sigma_n^2}$ where $\sigma_0=1.5\sqrt{2}$ so that G(x,y,1) = $G_gI(x,y)$. The remaining Gaussian images are created using $\sigma=1.5\left(\sqrt{2}\right)^m$ (m = 0,1,2,3) resulting in five Gaussian blurred images (G(x,y,s) (s = 0,…,4)). The size of the Gaussian filter is always the closest odd number to 3σ. These parameters were selected empirically and are the same for all images. Then the four DoG images are created by subtracting each Gaussian image from the previous Gaussian image in scale:

$$D(x,y,s)=G(x,y,s+1)-G(x,y,s) \quad (s=0,1,2,3) \quad (2)$$

For D(x,y,1) and D(x,y,2) the local minima and maxima with the highest magnitude are found in each region so that every region contains a potential feature point unless some portion of that region is occluded.

Feature point stability verification: To solve the problem of stability, Lowe fit a 3-D Quadratic function to the selected points. Using the Taylor Expansion of the DoG images D(x,y,s):

$$D(\tilde{x})=D+\frac{\partial D^T}{\partial \tilde{x}}\tilde{x}+\frac{1}{2}\tilde{x}^T\frac{\partial^2 D}{\partial \tilde{x}^2}\tilde{x} \quad (3)$$

Taking the derivative of this function with respect to $\tilde{x}$ and setting it equal to zero, we determine the extremum, $\tilde{x}$, to be

$$\tilde{x}=-\frac{\partial^2 D^{-1}}{\partial \tilde{x}^2}\frac{\partial D}{\partial \tilde{x}} \quad (4)$$

To reject points that have low contrast, we substitute Eq.(4) into Eq.(3) which results in

$$D(\tilde{x})=D+\frac{1}{2}\frac{\partial D^T}{\partial \tilde{x}}\tilde{x} \quad (5)$$

If $\left|D(\tilde{x})\right|$ is less than 0.03 (value from Lowe [26]) for a given extrema point, that point is rejected.

Dominant orientation: For each selected extrema point, the dominant orientation of the gradient of the points within a window around the feature point is determined from the smoothed histogram of orientations. The magnitude and orientation of the gradient of each point is found and the orientation is stored within one of36 bins. For each point within a window of W around the feature point, the weighted gradient magnitude is added to the bin corresponding to that point's orientation.

Feature description: Knowing the dominant orientation of each point, a descriptor can be created using the relative orientation and magnitude of the gradient and relative position of each point within a window, W, with respect to the feature point and its dominant orientation. The feature descriptor has a size of 64 bins: four bins for x direction, four bins for y direction and four bins for orientation. So, based on the normalized, relative orientation and position of each point in the normalized window with respect to the feature point.

The search window updating: The change of object location requires an efficient and adaptive search window updating mechanism. The proper search window location ensures that the object always lies within the search area and thus prevents loss of the object inside the search window. In this paper we use SIFT algorithm for localize the search window.

We Consider a points in the frame n with two component x_n^i and y_n^i. N_{Sift} is denoted to the number of key issues, identified using the above mentioned method. The key issue s are matched between the frame n-1 and frame n. the motion of the key point j from frame n-1 to frame n is shown by dx_n^j and dy_n^j. The estimated position of the object will be given by the sum of the position and the average motion estimated by sift algorithm:

$$x_{new}=\left(\frac{1}{N_{sift}}\sum_{j=1}^{N}dx_n^j\right)+x_{n-1}^i \quad (6)$$

$$y_{new}=\left(\frac{1}{N_{sift}}\sum_{j=1}^{N}dy_n^j\right)+y_{n-1}^i \quad (7)$$

Sequential monte carlo framework: The aim of sequential Monte Carlo estimation is to evaluate the *posterior* Probability Density Function (PDF) $p(X_k|Z^k)$ of the state vector X_k given a set $Z^k = \{z_1,\ldots,z_k\}$ of sensor measurements at a time k. The Monte Carlo approach relies on a sample-based construction to represent the state PDF. Multiple particles (samples) of the state are generated, each one associated with a weight which characterizes the quality of a specific particle. An estimate of the variable of interest is

obtained by the weighted sum of particles. Two major stages can be distinguished: *prediction* and *update*. During the prediction each particle is modified according to the state model of the region of interest in the video frame, including the addition of random noise in order to simulate the effect of the noise on the state. In the update stage, each particle's weight is re-evaluated based on the new data. A *resampling* procedure deals with the elimination of particles that have small weights and replication of the particles with larger weights.

Particle filtering: We denote by X_t a target state at time t, z_t the observation data at time t and $Z_t = \{z_1,\ldots,z_t\}$ all the observations up to time t. Assuming a non-Gaussian state space model, the prior probability $p(X_t|Z_{t-1})$ at time t in a Markov process is defined as:

$$p(x_t|Z_{t-1}) = \int p(X_t|X_{t-1})p(X_{t-1}|Z_{t-1})dx_{t-1} \quad (10)$$

where $p(X_t|Z_{t-1})$ is a state transition distribution and $p(X_{t-1}|Z_{t-1})$ stands for a posterior probability at time t- 1. The posterior probability which the tracking system aims to estimate at each time is defined as:

$$p(X_t|Z_t) \propto p(z_t|X_t)p(X_t|Z_{t-1}) \quad (11)$$

where $p(z_t|X_t)$ is the data likelihood at time t. According to the particle filtering framework, the posterior $p(X_t|Z_t)$ is approximated by a Dirac measure on a finite set of P particles $\{X_t^i\}_{i=1\ldots p}$ following a sequential Monte Carlo framework [27, 28]. Candidate particles are sampled by a proposal transition kernel $q(\tilde{X}_t^i|X_{t-1}^i, z_{t-1})$. The new filtering distribution is then approximated by a new sample set of particles $\{\tilde{X}_t^i\}_{i=1\ldots p}$ having the importance weights $\{w_t^i\}_{i=1\ldots p}$, where

$$w_t^i \propto \frac{p(z_t|\tilde{x}_t^i)p(\tilde{x}_t^i|x_{t-1}^i)}{q(\tilde{x}_t^i|x_{t-1}^i, z_{t-1})} \text{ and } \sum_{i=1}^{p} w_t^i = 1 \quad (12)$$

The sample set $\{X_t^i\}_{i=1\ldots p}$ can then be obtained by resampling $\{\tilde{X}_t^i\}_{i=1\ldots p}$ with respect to $\{w_t^i\}_{i=1\ldots p}$. By default, the Bootstrap filter is chosen as proposal distribution: $q(\tilde{X}_{t-1}^i, z_{t-1}) = p(\tilde{X}_t^i|X_{t-1}^i)$ Hence the weights can be computed by evaluating the corresponding data likelihood. We denote by D, the overall energy function where is energy related to texture cues. Thus, to favor candidate regions which FV distribution is similar to the reference model at time t.

Texture cue: A measure, *d*, is defined using the Bhattacharyya coefficient characterizing the difference (distance) between two normalized feature vectors $\bar{t}_1$ and $\bar{t}_2$.

$$d(\bar{t}_1, \bar{t}_2) = \sqrt{1 - \rho(\bar{t}_1, \bar{t}_2)} \quad (13)$$

where the Bhattacharyya coefficient, $\rho(\bar{t}_1, \bar{t}_2)$ is defined as:

$$\rho(\bar{t}_1, \bar{t}_2) = \sum_{i=1}^{m} \sqrt{\bar{t}_{1,i}\,\bar{t}_{2,i}} \quad (14)$$

The Texture Likelihood can then be defined as:

$$\zeta_{texture}(Z_{texture}|x) \alpha \exp\left(-d^2(\bar{t}_x, \bar{t}_{ref})/2\sigma_t^2\right) \quad (15)$$

where σ_t the standard deviation of the Gaussian texture noise is, $\bar{t}_x$ is the feature vector of the current frame and $\bar{t}_{ref}$ is the reference feature vector.

The larger the coefficient $\rho(\bar{t}_1, \bar{t}_2)$ is the more similar the distributions are. The Bhattacharyya distance values are within the interval [0,1]. For two identical feature vector we obtain d = 0 (ρ = 1) indicating a perfect match.

EXPERIMENTAL RESULTS

In this section, we performed several experiments to prove the feasibility of the proposed tracking method (Fig. 4). These sequences consist of indoors and outdoors testing environments so that the proposed scheme can be fully evaluated.

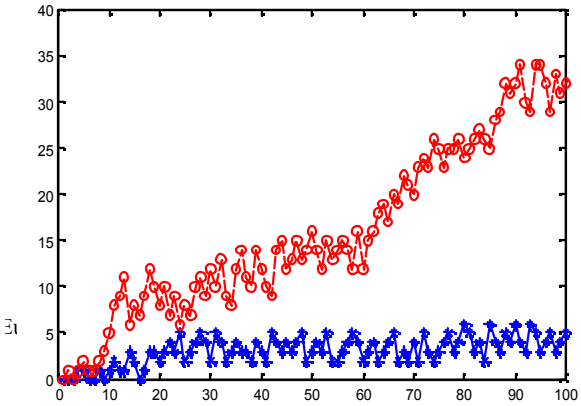
The experimental results of the proposed tracking algorithm have been compared with well-known color histogram based tracking algorithms with Bhattacharyya matching distance. We have used biorthogonal wavelet bases, which are particularly useful for object detection and generation of the UWPT tree. In the color histogram-based algorithm implementation, The RGB color space was taken as feature space and it was quantized into 16*16*16 bins. The Epanechnikov profile was used for histogram computations. It must be pointed out that in this evaluation there is no intension to track multiple objects. On the contrary, a single object is detected in the first frame of each sequence, followed by continuous tracking to the remaining part of the sequence. In some sequences, there is more than one object in the scene.

Fig. 4: Test sequences used in current evaluation

(a) Reference: frame no. 1

(b) UWPT: frame no. 24 UWPT: frame no. 52 UWPT: frame no. 70 UWPT: frame no. 84

(c) CHB: Frame no. 24 CHB: Frame no. 52 CHB: Frame no. 70 CHB: Frame no. 84



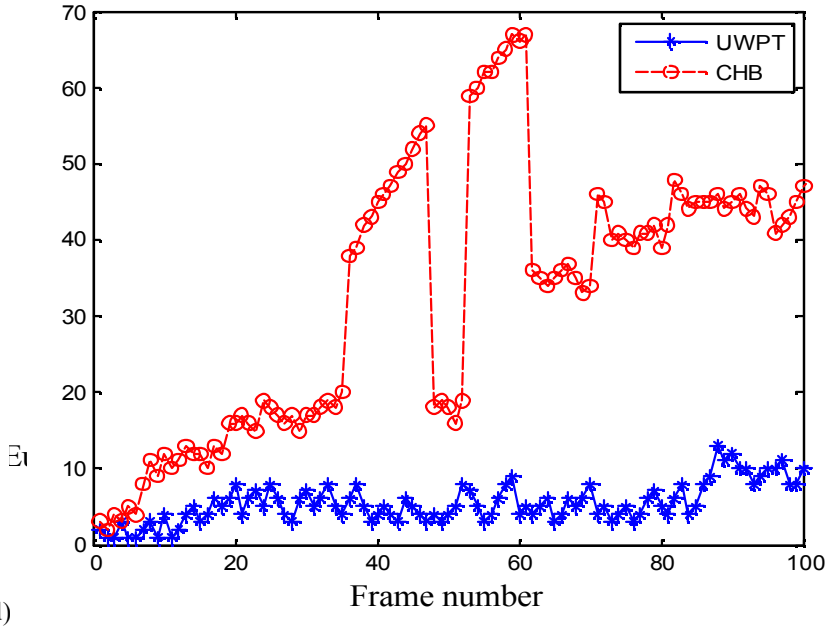
(d)

Fig. 5: Tracking the head of a man coming down the stairs in a crowded metro station. (a) Reference frame, (b) UWPT, (c) CHB (d) Objective evaluation: distance between the center of tracked bounding box and the expected center, for all methods

(a) Reference: frame no. 1

(b) UWPT: Frame no. 23 UWPT: Frame no. 39 UWPT: Frame no. 51 UWPT: Frame no. 65

(c) CHB: Frame no. 23 CHB: Frame no. 39 CHB: Frame no. 51 CHB: Frame no. 65



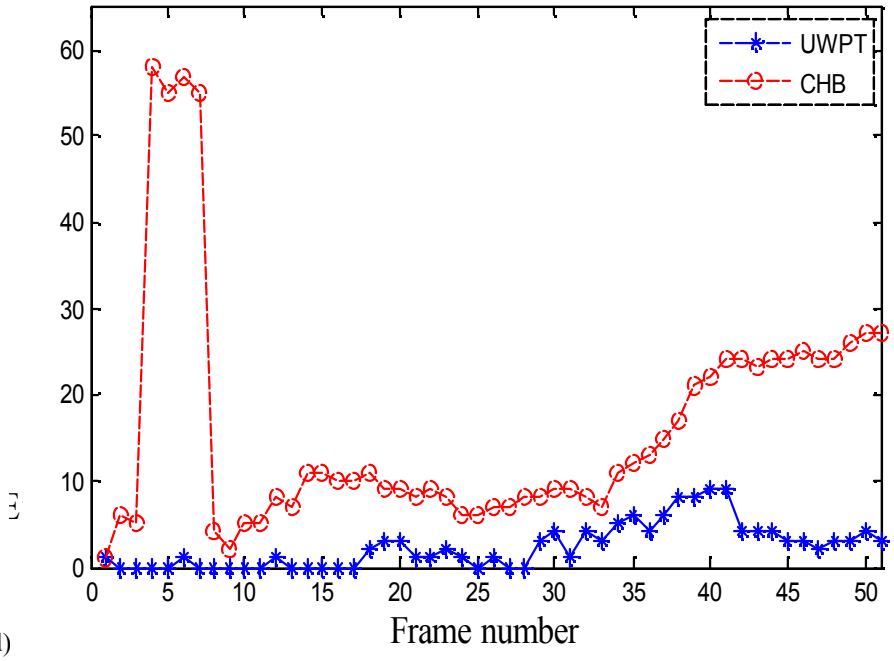
(d)

Fig. 6: Tracking the head of a man coming down the stairs in a crowded metro station in presence of additive Gaussian white noise (PSNR =20 dB). (a) Reference frame, (b) UWPT, (c) CHB (d) Objective evaluation: distance between the center of tracked bounding box and the expected center, for all methods

(a) Reference: frame no. 11

(b) UWPT: Frame no. 18 UWPT: Frame no. 29 UWPT: Frame no. 53 UWPT: Frame no. 63

(c) CHB: Frame no. 18 CHB: Frame no. 29 CHB: Frame no. 53 CHB: Frame no. 63



(d)

Fig. 7: Tracking the head of a man coming up the stairs in a crowded metro station. (a) Reference frame, (b) UWPT, (c) CHB (d) Objective evaluation: distance between the center of tracked bounding box and the expected center, for all methods

Reference: frame no. 1

b) UWPT: Frame no. 6 UWPT: Frame no. 45 UWPT: Frame no. 77 UWPT: Frame no. 108

c) CHB: Frame no. 6 CHB: Frame no. 45 CHB: Frame no. 77 CHB: Frame no. 108

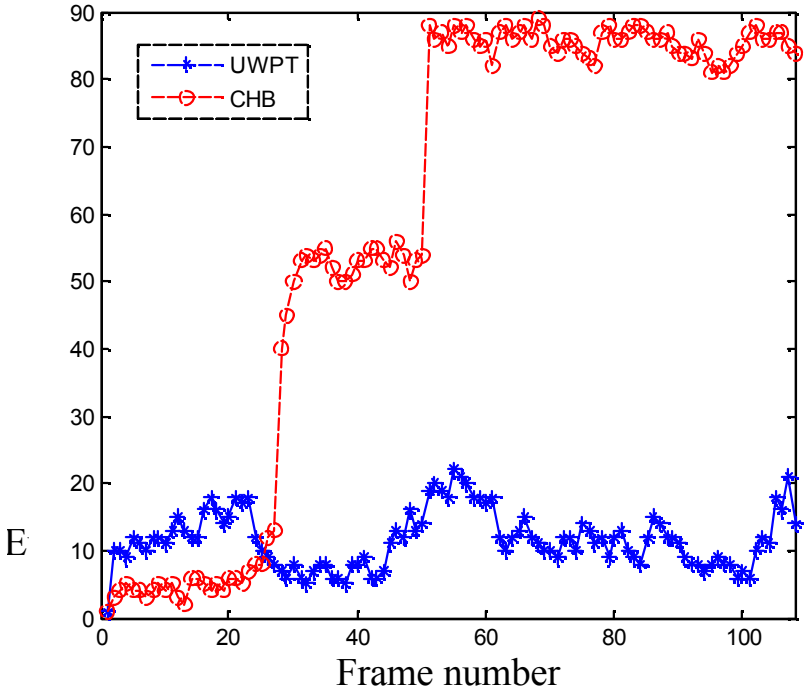


Fig. 8: Tracking the a man. (a) Reference frame, (b) UWPT, (c) CHB and (d) Objective evaluation: distance between the center of tracked bounding box and the expected center, for all methods

To evaluate the algorithms in a real-environment setting, we have applied them to different real-time video clips of Tehran Metro Stations. These video clips show the crowds at different parts of the metro such as getting on/of the train and up/down the stairs. Moreover, they include different conditions in crowded scenes such as partial, high and low speed, object deformation and object rotation. Note the difficulty in tracking heads in a crowded scene, as there are several nearby similar objects. We have defined a measure for objective evaluation of tracking techniques based on the Euclidian distance of the center of gravity of the tracked and actual objects. Here, at the start of tracking, a bounding rectangle located at the center of the gravity of the desired object is selected. In the following frames, the bounding rectangle represents the tracked object and its distance with the center of the gravity of the actual object is measured.

Figure 5 shows the snapshots of tracked head of a man, shown in frame 24, coming down the stairs in a crowded metro station. The object is stepping down the stairs with a constant speed, small amount of zooming and some cross-movements. There is no partial or full occlusion of the object in this case, but there are similar faces that complicate the tracking process. As the results show, the object of interest has been successfully tracked by UWPT despite the presence of several similar objects. For color histogram based tracking method, their Euclidian distance of the center of gravity is well above the bounding rectangle size, implying that the objects are totally miss-tracked.

Figure 6 shows the result of tracking the sequence in the presence of additive white Gaussian noise with a peak Signal-to-noise ratio (PSNR) of 20 dB. This type of noise is very common with low-light video, especially in undergrounds. Again, the noisy reference frame is frame no. 1 (Fig. 6(a)).

The presence of noise has degraded the performance of the color histogram-based algorithms tremendously without having any effect on the wavelet-based methods (Fig. 6(f)).

Figure 7 shows the result of tracking a person moving up the stairs and away from the camera in a metro station. Frame no. 11 was the reference frame (Fig. 7(a)). The target object is stepping up the stairs with a constant speed and its movement exhibits a small amount of zooming out, some degree of rotation of the head and partial occlusion.. In all the sequences, our proposed algorithm can successfully track the target object even in the presence of object rotation and partial occlusion.

CONCLUSION

Many of the existing algorithms for object tracking that are based on spatial domain features; fail in the presence of change in appearance or pose or in the presence of noise. The tracking framework necessitates robust and efficient but accurate methods for segmentation and matching. To overcome these problems, in this paper, we have proposed a new method of object tracking using structural similarity index in wavelet transform domain and sift feature, which is approximately shift-invariant. Sift features are used to correspond the region of interests across frames in search window. The sift features are invariant to image scaling, translation and rotation and partially invariant to illumination changes and affine or 3D projection. The reference object in the initial frame is modeled by a feature vector in terms of the coefficients of Undecimated wavelet transform. A similarity measure based on structural similarity index is used to find the object in the current frame. The particle filter maintains the estimation of the object positions and it makes predictions about its movement. This makes the system robust to temporal occlusions of the object. The proposed tracking algorithm yields better results even in noisy video as shown in the experiments.

ACKNOWLEDGMENTS

This work has been conducted with the UK MOD Data and Advanced Information and Communication Technology Center (AICTC) of Sharif University of Technology.

REFERENCES

1. Liu, T.-L. and H.-T. Chen, 2004. Real-time tracking using trust region methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26 (3): 397-402.
2. Welch, G. and G. Bishop, A introduction to the Kalman fitler. Technical report TR 95-041.
3. Ristic, B., S. Arulampalam and N. Gordon, 2004. Beyond the Kalman filter: Particle filters for tracking applications, Artech House.
4. Nummiaro, K., E. Koller-Meier and L. Van-Gool, 2003. An adaptive color-based particle filter. Image and Vision Computing, 21: 99-110.
5. Rasmussen, C. and G. Hager, 2001. Probabilistic data association methods for tracking complex visual objects. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23 (6): 560-576.
6. Cham, T. and J. Rehg, 1999. A Multiple Hypothesis Approach to Figure Tracking. Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2: 239-219.

7. Comaniciu, D., V. Ramesh and P. Meer, 2003. Kernel-based object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25 (5): 564-577.
8. Xu, D., Y. Wang and J. An, 2005. Applying a new spatial color histogram in mean-shift based tracking algorithm. In Proceedings of the Image and Vision Computing New Zealand (IVCNZ '05), University of Otago, Dunedin, New Zealand.
9. Lucena, M., J.M. Fuertes and N.P. de la Blanca, 2004. Evaluation of three optical flow based observation models for tracking. ICPR, pp: 236-239.
10. Wu, Y. and T.S. Huang, 2000. A co-Interface approach to Robust Visual Traking. Proc-intl conf. Computer vision and pattern recognition, 2: 134-141.
11. Rabiner, L., 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proc. IEEE 77, pp: 257-286.
12. Barlow, H., 1972. Single units and sensation: A neuron doctrine for perceptual psychology, Perception, 1: 371-394.
13. Ullman, S., 1979. The Interpretation of Visual Motion. MIT Press, Cambridge, MA.
14. Cheng, F.-H. and Y.-L. Chen, 2006. Real timemultiple objects tracking and identification based on discrete wavelet transform. Pattern Recognition, 39 (6): 1126-1139.
15. Coifman, R.R. and M.V. Wickerhauser, 1992. Entropy-based algorithms for best basis selection. IEEE Transactions on Information Theory, 38 (2): 713-718.
16. Guo, H., 1995. Theory and applications of shift invariant, timevarying and undecimated wavelet transform. M.S. Thesis, Rice University, Houston, Tex, USA.
17. Khansari, M., H.R. Rabiee, M. Asadi, M. Ghanbari, M. Nosrati and M. Amiri, 2005. A quantization noise robust object's shape prediction algorithm. In Proceedings of the 13th European Signal Processing Conference (EUSIPCO '05), Antalya, Turkey.
18. Khansari, M., H.R. Rabiee, M. Asadi, M. Nosrati, M. Amiri and M. Ghanbari, 2006. Object shape prediction in noisy video based on undecimated wavelet packet features. In Proceedings of the 12th International Multimedia Modelling Conference (MMM '06), Beijing, China.
19. Jinqyuan, Z., J. Xingzhou and Y. Bingcheng, 1997. Adaptive wavelets classification of transient sonar signals. International Conference on Signal Processing, Vol: 6 (5).
20. Strickland, R.N. and H.I. Hahn, 1997. Wavelet transform methods for object detection and recovery. IEEE Transaction on Image Processing, Vol: 6 (5).
21. Messer, K., D. de Ridder and J. Kittler, 1999. Adaptive texture representation methods for automatic target recognition. Proc. of British Machine Vision Conference, BMVC-9, Nottingham, UK.
22. Coifman, R.R. and M.V. Wickerhauser, 1992. Entropy-based algorithms for best basis selection. IEEE Transaction on Information Theory. Special Issue on Wavelet Transforms and Multiresolution Signal Analysis, Vol: 38 (5).
23. Zhang, H., W. Gao, X. Chen and D. Zhao, 2005. Learning informative features for spatial histogram-based object detection. In Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN '05). Montreal, Quebec, Canada, 3: 1806-1811.
24. Khansari, M., H.R. Rabiee, M. Asadi, M. Ghanbari, M. Nosrati and M. Amiri, 2005. A shape tracking algorithm based on generated pixel features by undecimated wavelet packet. In Proceeding of the 10th Annual Computer Society of Iran Computer Conference (CSICC '05), Tehran, Iran.
25. Khansari, M., H.R. Rabiee, M. Asadi, P. Khadem Hamedani and M. Ghanbari, 2006. Adaptive search window for object tracking in the crowds using undecimated wavelet packet features. In Proceedings of the World Automation Congress (WAC '06), Budapest, Hungary, pp: 1-6.
26. Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision, 60 (2): 91-110.
27. Arulampalam, M., S. Maskell, N. Gordon and T. Clapp, 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. on Signal Proc., 50 (2): 174-188.
28. Hue, C., J.-P.L. Cadre and P. Pe´rez, 2002. Tracking multiple objects with particle filtering. IEEE Transactions on Aerospace and Electronic Systems, 38 (32): 791-812.