Military Technical College
Kobry El-Kobbah,
Cairo, Egypt

11th International Conference
on Electrical Engineering
ICEENG 2018

DUAL-TREE COMPLEX WAVELET TRANSFORM FOR ROBUST VISUAL TRACKING OF MULTIPLE OBJECTS IN VIDEO

Ahmed Mahmoud * and Sherif S. Sherif **

ABSTRACT

We developed a robust multiscale visual tracker of multiple objects in video using the *Dual-tree Complex Wavelet Transform* (DT-CWT). Real-valued wavelet transforms were previously used for visual tracking, but most suffer from *shift variance* and lack of *directional selectivity*. Therefore, we used DT-CWT to avoid such shortcomings. In our tracker, a captured video frame was represented as different subbands using DT-CWT. Then we applied N independent particle filters to a small subset of these subbands, where the choice of these subbands changed adaptively with each captured frame. Finally, we fused the position tracks resulting from these particle filters to obtain final position tracks of multiple moving objects in the video. To demonstrate robustness of our visual tracker, we compared the performance of our multiscale tracker to a standard particle filter full resolution-based tracker and a single wavelet subband $(LL)_2$ based tracker, our multiscale tracker demonstrates significantly better tracking performance.

KEY WORDS
Robust visual tracking, particle filter, dual-tree complex wavelet transform.

1- INTRODUCTION

Visual tracking is an important research problem because of its broad variety applications that include biomedical, industrial and security applications. Visual tracking aims to estimate the states of a single or a group of moving object(s), e.g., position and velocity, in a video sequence. Over the past few decades, significant progress was made on visual tracking, however robust visual tracking still remains an active research topic [1, 2]. Robust visual tracking relates to the capability to avoid tracking failures [3] and to track objects accurately in video sequences that have challenging conditions and unexpected events [4]. These challenging conditions and unexpected events could comprise the presence of 1) background motion and object shadows; 2) objects with different sizes and contrast level; 3) low signal-to-noise (SNR) ratio; 4) sudden change in scene illumination; and 5) partial object camouflage.

* Egyptian Armed Forces.
** Dept. of Electrical and Computer Engineering, University of Manitoba, Canada.

Domain transformation could be a possible approach for robust visual tracking, as the representation of video sequences in a different domain can suppress effects of both noise and sudden changes in illumination. To handle the presence of background motion and changes in illumination, the work in [5] used a two-dimensional discrete wavelet transform (DWT) to represent a video sequence, but this tracker used only the low-frequency subband. As robust tracking of objects with different contrasts and different sizes would require information from all DWT scales [6], robust tracking of such challenging objects would be likely unachievable using this method.

To detect the edges map of a moving object at time t, another work in [7] used the dual-tree complex wavelet transform (DT-CWT) to represent three consecutive video frames (Z_{t-1}, Z_t, and Z_{t+1}). Then two sets of difference-frames were computed based on the difference between the corresponding subbands at time $t-1$ and t, and the corresponding subbands at time t and $t+1$. Then a merger between corresponding difference-frames of the two sets were used to obtain the final edge map of the moving object.

Another possible approach for robust visual tracking is to use data fusion in a sequential Bayesian framework [4, 8]. Fusion is a common approach to improve the accuracy and robustness of a visual tracker [9], where it could be performed by fusing 1) multiple visual features (cues) in a video frame, 2) independent sources of measurements, or 3) tracking paths from different visual trackers, i.e., tracker-level fusion. We note that sequential Bayesian trackers, e.g., Kalman filters or particle filters could be viewed as information fusers [10] due to their ability to combine observation data and a dynamic model for the object into one mathematical framework. That is why they are attractive, even though they could be computationally expensive especially as the number of tracked objects increases [11].

In this paper, we describe a robust multiscale visual tracker that represents a captured video frame as different subbands of the dual-tree complex wavelet transform. It then applies N independent particle filters to a small subset of these subbands, where the choice of this subset of wavelet subbands changes adaptively with each captured frame. Finally, it fuses the outputs of these N independent particle filters, i.e., tracker-level fusion, to obtain the final position tracks of multiple moving objects in the video sequence.

2- BAYESIAN VISUAL TRACKING

A conventional Bayesian method for visual tracking uses standard particle filter applied to the full resolution video frame. The particle filter provides an approximate solution to the states of the moving objects by representing a point mass function as a weighted sum of random samples that are usually called particles [12].
This conventional method could have many limitations when tracking multiple objects in video sequences that have challenging visual conditions and include unexpected events, including:

1. Presence background motion and shadow could create additional (spurious) likelihood modes.

2. Presence of objects with different sizes and contrast levels could lead to a dominant likelihood problem [13, 14], in which the particle filter's posterior distribution contains multi-modes that represent different objects, but it would be dominated by a single object (likelihood mode) that has the largest size and/or highest intensity.
3. Presence of a high level of noise in the video frames could create additional (spurious) likelihood modes,
4. Presence of sudden changes in illumination could lead to sudden changes in the likelihood function
5. Presence of partial object camouflage could produce sudden changes in the likelihood modes

3- DUAL-TREE COMPLEX WAVELETS

The dual-tree complex wavelet transform was introduced by Nick Kingsbury [15] as an enhancement to the real-valued wavelet transform [16]. Compared to the real-valued wavelet transform, the DT-CWT has additional superior features, including:

1. More ***directional selectivity*** feature, i.e., it detects the edges in an image along six directions at different resolution scales, compared to the DWT that detects edges at the vertical, horizontal and diagonal orientations only.
2. ***Shift invariance*** feature, i.e., a shift of a signal does not produce a shift in the coefficients of the subbands. This is achieved with a limited redundancy factor of only 4 for 2-D images [16].

To address the above limitations of using a standard particle filter for visual tracking, we used the *dual-tree complex wavelet transform* to represent captured video frames, as it has several advantages for visual tracking, including:

1. A wavelet transform, e.g., the *dual-tree complex wavelet transform*, would be suitable for tracking objects with different sizes and/or contrast levels that could be present in the same video frame, as it produces subband frames having different resolutions (scales). Subband frames with a coarse resolution (large scale) are more suitable for tracking large objects and/or objects with a high contrast, while subband frames with a fine resolution (small scale) are more appropriate for tracking small objects and/or objects with a low contrast [6].
2. In visual tracking, different types of object motion, *e.g.,* translation and rotation are detected across subsequent frames. Therefore, representing these motion translations using a shift invariant transform, e.g., *Dual-tree complex wavelet transform* could produce better visual tracking results [17].
3. *Dual-tree complex wavelet transform* is a natural edge detector that could detect boundaries of objects in various directions. It is sensitive to edges along six different directions ($\pm 15^0,\ \pm 45^0, \text{and}\ \pm 75^0$).
4. Denoising of a video sequence using wavelets is relatively easy. Typically, it is performed by setting small wavelet coefficients to zero. Compared to using a shift variant wavelet transform, using a shift invariant wavelet transform, e.g., the *dual-tree complex wavelet transform,* would typically result in better denoising performance [18].

4- IMPLEMENTATION OF VISUAL TRACKER

To start visual tracking, we constructed a *background frame* from the full-resolution video sequence. Then we applied the *dual-tree complex wavelet transform* to both background and current frames to generate the subband frames. Then we subtracted the subbands of the background frame from their corresponding subbands of the current frame to produce subband *difference frames*. We then applied three independent particle filters to three adaptively chosen subbands of the difference frame. We obtained our final position tracks by fusing the position tracks that resulted from our three subband particle filters.

5- Background Extraction and Update

We detected moving objects in the video sequence by constructing a frame that represents the background. We chose the Long-Term Average Background Modeling (LTABM) background extraction method [19], as it is a fast technique that suits the real-time requirement of visual tracking. We constructed the initial background frame, $\boldsymbol{B_0}$, by averaging the first few frames of the video sequence. Then we transformed, $\boldsymbol{B_0}$, to the complex wavelet domain as described in Section 4.2, to obtain, $\boldsymbol{B_0^s}$, the initial subband background frames at different $\boldsymbol{s}$ scales. At every time instance, the subband background frames, $\boldsymbol{B_t^s}$, were updated as described in [19].

6- Generation of Subband Frames

We generated the subband frames, Z_t^s, at different scales *s* using the *dual-tree complex wavelet transform*. This transform produced two low-frequency subbands, and six high-frequency subbands that represent detected edges at various orientations. The scale s belongs to the set of all subband frames in levels one and two of the complex wavelet tree.

7- Generation and Selection of Subband Difference Frames

We generated subband *difference frames*, D_t^s, by subtracting the current background B_t^s from the current frame Z_t^s. As an approximation of the energy in a subband, we calculated the l_1 norm for each subband *difference-frame.* Then we kept the three subbands having the highest l_1 norm values and discarded the rest. We note that discarding the subbands having the lowest energies would result in denoising.

8- Generation of Subband Binary Frames and Labeling of Objects

We generated subband binary frames from the three chosen subband difference frames through thresholding. Pixels with values above a positive threshold were categorized as foreground, and the resulting white pixels in these subband binary frames would represent candidate moving objects. After generating these binary frames, we implemented morphological operations, including dilation and fill operations, to enhance the shapes of the present objects. We then obtained subband

labeled frames by identifying the present connected pixel regions, through scanning the subband binary frames pixel by pixel from left to right and top to bottom.

Implementation of our Subband Particle Filters

We implemented three independent particle filters, where each filter processed one of the three subband labeled binary frames obtained in the previous section. These subband particle filters continuously updated the kinematic states of the objects present in these labeled binary frames. We note that we used a linear motion model similar to the one described in [20], and a measurement model based on motion cues analogous to the one described in [21].

9- Fusion of the Resulting Position Tracks

Our subband particle filters produced three sets of position tracks corresponding to potential moving objects. To obtain the final set of position tracks, we performed an object confirmation step by performing a voting for the presence of an object in a predefined area, followed by an averaging of the position tracks of confirmed objects. To associate an object i at time t with an object j at time $t-1$, we performed one or possibly two inter-frame data association steps. First we used position-gating method described in [22] which imposes a distance constrain to associated object i with an object, j. If position-gating method failed, we would resort to gray-scale histogram comparison that is similar to the one described in [21].

10- EVALUATION OF TRACKING PERFORMANCE

To demonstrate the improved performance of our multiscale tracker, compared to a typical visual tracker using a standard full-resolution particle filter, and to a single wavelet subband $(LL)_2$ based tracker [23], we applied it to a challenging video sequence that included background motion, shadow, and partial object camouflage.

11- Example Demonstrating Object Shadow and Partial Object Camouflage

The video sequence in this example, *"OneLeaveShopReenter2front"* is from the Caviar database (288 x 384 pixels, 25 *fps*, 558 frames). In this video sequence, two people walk in front of a store, while another person exits the store and then re-enters it. To quantitatively compare the performance of these three visual trackers, we will define a *detection frame* of a specific object as a frame where this particular object was correctly detected by these three trackers. As shown in Table 1, object 1 in this video appeared in 54 *detection frames*, with cumulative track errors of 303 pixels, 439 pixels, 362 pixels using 1) standard full resolution particle filter, 2) single wavelet subband $(LL)_2$, and 3) our multiscale tracker, respectively. Object 2 in this video appeared in 456 detection frames, with cumulative track errors of 1820 pixels, 3296 pixels, 2829 pixels using 1) standard full resolution particle filter, 2) single wavelet subband $(LL)_2$, and 3) our multiscale tracker, respectively. Object 3 in this video appeared in 116 *detection frames*, with cumulative track errors of 1518 pixels, 1442 pixels, 632 pixels using 1)

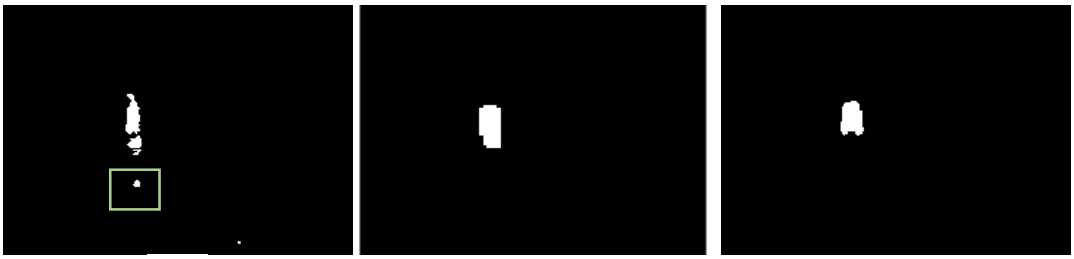
standard full resolution particle filter, 2) single wavelet subband $(LL)_2$, and 3) our multiscale tracker, respectively. These values are a solid demonstration of the superior performance and robustness of our multiscale tracker compared to the other two trackers.

Table 1. Number of missed object events, average position track errors, and number of phantom object events

Visual tracker type	Missed object (event/558 frames)	Average position track error (pixel/*detection frame*)			Phantom object (event/558 frames)
		Object 1	Object 2	Object 3	
Full resolution particle filter tracker	55	5.62	3.99	13.08	469
$(LL)_2$ subband tracker	80	8.13	7.23	12.43	2
Our multiscale tracker	33	6.7	6.2	5.45	1

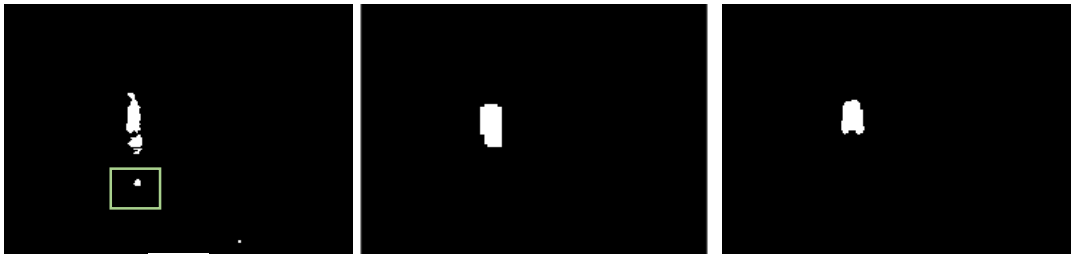
12- Demonstrating challenging video conditions

12.1 Object shadow:



(a) (b) (c)

Fig. **1** (a) highlights an artifact due to object shadow. Fig. 2 (a), Fig. 2 (b), and Fig. 2 (c) the visual tracking results, superposed on the 305th video frame, generated by the standard full-resolution particle filter-based tracker, the single wavelet subband $(LL)_2$ based tracker, and our DT-CWT based visual tracker, respectively.



(a) (b) (c)

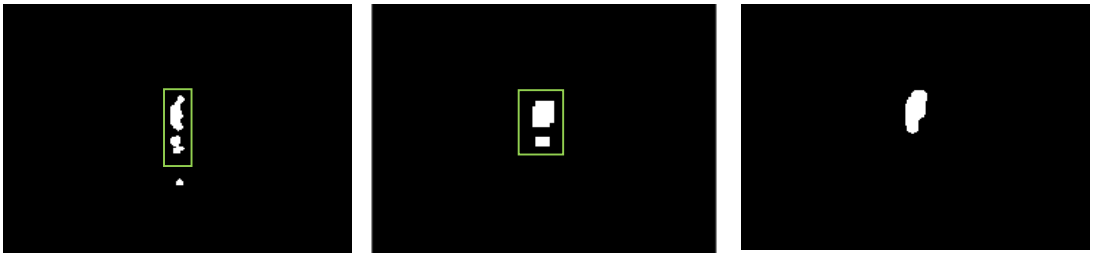
Fig. 1. Binary frames generated from the 305th frame using: (a) the full resolution frame; (b) subband $(LL)_2$; (c) a chosen subbands in our multi-scale tracker

(a) (b) (c)

Fig. 2. Visual tracking results for the 305th video frame using: (a) the standard full-resolution particle filter-based tracker; (b) our multi-scale DT-CWT based tracker

We note that our multi-scale tracker overcame the presence of object shadow in this video frame.

12-Partial object camouflage: Fig. 3 (a), Fig. 3 (b), and Fig. 3 (c) show the binary frames generated from the 427th video frame using the full-resolution frame, subband $(LL)_2$, and, one of chosen subbands in our multi-scale DT-CWT based tracker, respectively. We note that the green box in Fig. 3 (a) and Fig. 3 (b) highlights the division of an object into two due to partial object camouflage. Fig. 4 (a), Fig. 4 (b), and Fig. 4 (c) show the visual tracking results, superposed onto the 427th video frame, generated by the standard full-resolution particle filter-based tracker and our multi-scale DT-CWT based tracker, respectively. We note that our visual tracker overcame the presence of partial object camouflage in this video frame.



(a) (b) (c)

Fig. 3. Binary frames generated from the 427th frame using: (a) the full resolution frame; (b) subband $(LL)_2$; (c) a chosen subband by multi-scale tracker

(a) (b) (c)

Fig. 4. Visual tracking results for the 427th video frame using: (a) the standard full-resolution particle filter-based tracker; (b) our multi-scale DT-CWT based tracker

13- Example Demonstrating a Change in Illumination and Objects of Different Sizes

In the second example, we used the *"Meet_WalkTogether2"* video sequence from the CAVIAR database. In this sequence, two people meet and walk together. The challenging conditions present in this video sequence are a change in illumination and the presence of objects of different sizes. To quantitatively compare performance of the three visual trackers considered here, Table 2. shows that object 1 in this video appeared in 109 *detection frames*, with cumulative track errors of 721 pixels, 738 pixels, 547 pixels using 1) standard full resolution particle filter, 2) single wavelet subband $(LL)_2$, and 3) our multiscale tracker, respectively. Object 2 in this video appeared in 8 *detection frames*, with cumulative track errors of 72 pixels, 101 pixels, 80 pixels using 1) standard full resolution particle filter, 2) single wavelet subband $(LL)_2$, and 3) our multiscale tracker, respectively. Object 3 in this video appeared in 60 *detection frames*, with cumulative track errors of 718 pixels, 1023 pixels, 653 pixels using 1) standard full resolution particle filter, 2) single wavelet subband $(LL)_2$, and 3) our multiscale tracker, respectively. These values are a solid demonstration of the superior performance and robustness of our multiscale tracker compared to the other two trackers.

Table 2. Number of missed object events, average position track errors, and number of phantom object events

Visual tracker type	Missed object (event/827 frames)	Average position track error (pixel/*detection frame*)			Phantom object (event/827 frames)
		Object 1	Object 2	Object 3	
Full resolution particle filter tracker	81	6.4	9.04	12	122
$(LL)_2$ subband tracker	62	6.6	12.7	17.05	0
Our multiscale tracker	33	6	6.27	6.15	1

14- Demonstrating challenging video conditions

Presence of illumination change: Fig. 5 (a), Fig. 5 (b), and Fig. 5 (c) depict the binary frames generated from the 31st video frame using the full-resolution frame, subband $(LL)_2$, and one of chosen subbands in our multi-scale DT-CWT based tracker, respectively. We note that the green box in Fig. 5 (a) highlights an artifact due to the illumination change in the video frame.

(a) (b) (c)

Fig. 5. Binary frames generated from the 67th frame using: (a) the full resolution frame; (b) subband $(LL)_2$; (c) a chosen subband by multi-scale tracker

(a) (b) (c)

Fig. 6. Visual tracking results for the 67th video frame using: (a) the standard full-resolution particle filter-based tracker; (b) our multi-scale DT-CWT based tracker

Fig. 6 (a), Fig. 6 (b), and Fig. 6 (c) show visual tracking results, superposed on the 67th video frame, generated by the standard full-resolution particle filter-based tracker, the single wavelet subband $(LL)_2$ based tracker, and our multi-scale DT-CWT based tracker.We note our multi-scale tracker overcame the effect of sudden illumination change in this 67th video frame.

Objects of different sizes: Fig. 7 (a), Fig. 7 (b), and Fig. 7 (c) show the binary frames generated from the 200th video frame using the full-resolution frame, subband $(LL)_2$, and one of chosen subbands in our multi-scale DT-CWT based tracker, respectively. We note that the object sizes in Fig. 7 (c) are closer to each other than the object sizes in (a). Fig. 8 (a), Fig. 8 (b), and Fig. 8 (c) show visual tracking results, superposed on the 200th video frame, generated by the standard full-resolution particle filter-based tracker, the single wavelet subband $(LL)_2$ based tracker, and our multi-scale DT-CWT based tracker, respectively. We note that, due to the presence of a large object, the standard full resolution particle filter-based tracker failed to track the smaller object. Also, the single wavelet subband $(LL)_2$ based tracker failed to track the small object due to using only one subband in a fixed scale: the second scale. Conversely, our multi-scale DT-CWT based tracker was able to overcome these problems and successfully tracked both the large and small objects.

(a) (b) (c)

Fig. 7. Binary frames generated from the 200th frame using: (a) the full resolution frame; (b) subband $(LL)_2$; (c) a chosen subband by multi-scale tracker

(a) (b) (c)

Fig. 8. Visual tracking results for the 200th video frame using: (a) the standard full-resolution particle filter-based tracker; (b) our multi-scale DT-CWT based tracker

15- CONCLUSIONS

We developed a robust multiscale visual tracker of multiple objects in video using the Dual-tree Complex Wavelet Transform. A captured video frame was represented as different subbands using DT-CWT, then we applied N independent particle filters to a small subset of these subbands, where the choice of these subbands changed

adaptively with each captured frame. Finally, we fused the position tracks resulting from these particle filters to obtain final position tracks of multiple moving objects in the video. To demonstrate robustness of our visual tracker, we applied it to videos with challenging visual conditions. On comparing the performance of our multiscale tracker to a standard particle filter full resolution-based tracker, our tracker achieved significantly more accurate tracking results.

REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *Acm computing surveys (CSUR),* vol. 38, p. 13, 2006.

[2] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," *Neurocomputing,* vol. 74, pp. 3823-3831, 2011.

[3] T. A. Biresaw, A. Cavallaro, and C. S. Regazzoni, "Tracker-level fusion for robust bayesian visual tracking," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 25, pp. 776-789, 2015.

[4] N. Zheng and J. Xue, *Statistical learning and pattern analysis for image and video processing*: Springer Science & Business Media, 2009.

[5] F.-H. Cheng and Y.-L. Chen, "Real time multiple objects tracking and identification based on discrete wavelet transform," *Pattern Recognition,* vol. 39, pp. 1126-1139, 2006.

[6] R. Gonzalez and R. Woods, "Digital image processing: Pearson prentice hall," *Upper Saddle River, NJ,* 2008.

[7] T. Celik and K.-K. Ma, "Moving video object edge detection using complex wavelets," *Advances in Multimedia Information Processing-PCM 2008,* pp. 259-268, 2008.

[8] Y. Bar-Shalom and X.-R. Li, "Multitarget-multisensor tracking: principles and techniques," *Storrs, CT: University of Connecticut, 1995.,* 1995.

[9] E. Maggio and A. Cavallaro, *Video tracking: theory and practice*: John Wiley & Sons, 2011.

[10] J. R. Raol, *Data Fusion Mathematics: Theory and Practice*: CRC Press, 2015.

[11] G. M. Rao and C. Satyanarayana, "Visual object target tracking using particle filter: a survey," *International Journal of Image, Graphics and Signal Processing,* vol. 5, p. 1250, 2013.

[12] W. L. Dunn and J. K. Shultis, *Exploring Monte Carlo Methods*: Elsevier, 2011.

[13] Z. Khan, T. Balch, and F. Dellaert, "An MCMC-based particle filter for tracking multiple interacting targets," in *Computer Vision-ECCV 2004*, ed: Springer, 2004, pp. 279-290.

[14] H. Tao, H. S. Sawhney, and R. Kumar, "A sampling algorithm for tracking multiple objects," in *International Workshop on Vision Algorithms*, pp. 53-68.

[15] N. Kingsbury, "The dual-tree complex wavelet transform: a new efficient tool for image restoration and enhancement," in *Signal Processing Conference (EUSIPCO 1998), 9th European*, 1998, pp. 1-4.

[16] F. Jin, "Wavelet-based image and video processing," University of Waterloo, 2004.

[17] R. Singh, R. K. Purwar, and N. Rajpal, "A better approach for object tracking using dual-tree complex wavelet transform," in *Image Information Processing (ICIIP), 2011 International Conference on*, 2011, pp. 1-5.

[18] J.-L. Starck, F. Murtagh, and J. M. Fadili, *Sparse image and signal processing: wavelets, curvelets, morphological diversity*: Cambridge university press, 2010.

[19] H. Hassanpour, M. Sedighi, and A. R. Manashty, "Video frame's background modeling: Reviewing the techniques," *Journal of Signal and Information Processing,* vol. 2, p. 72, 2011.

[20] D. Rowe, I. Huerta, J. Gonzàlez, and J. J. Villanueva, "Robust multiple-people tracking using colour-based particle filters," in *Iberian Conference on Pattern Recognition and Image Analysis*, pp. 113-120.

[21] J. J. Pantrigo, J. Hernández, and A. Sánchez, "Multiple and variable target visual tracking for video-surveillance applications," *Pattern Recognition Letters,* vol. 31, pp. 1577-1590, 2010.

[22] A. Amditis, G. Thomaidis, G. Karaseitanidis, P. Lytrivis, and P. Maroudis, *Multiple Hypothesis Tracking Implementation*: INTECH Open Access Publisher, 2012.

[23] C.-H. Hsia, J.-S. Chiang, and J.-M. Guo, *Multiple Moving Objects Detection and Tracking Using Discrete Wavelet Transform*: INTECH Open Access Publisher, 2011.