A Novel Approach for People Counting and Tracking from Crowd Video

M. Ayyüce Kızrak Sağun
Electric-Electronics Eng. Dpt.
Halic University
Istanbul, Turkey
ayyucekizrak@halic.edu.tr

Bülent Bolat
Electronics &Communications Eng. Dpt.
Yildiz Technical University
Istanbul, Turkey
bbolat@yildiz.edu.tr

***Abstract*— Crowd analysis on video recordings is an important research area currently. In this work, a combined crowd density estimation method is presented to overcome this problem. To improve the accuracy of the system two different estimators run simultaneously and a blob is marked as a person only if both estimators mark it as person. One of the main problems in crowd density estimation is occlusion. To overcome this problem we tracked the trajectories of blobs by using a Kalman filter. The method was applied to three common benchmark data which are PETS2009, UCSD and Grand Central. The results confirm the proposed method's success.**

Keywords—Crowd density; SIFT; complex wavelet transform;Optical flow; Kalman filtering; Video processing;

I. INTRODUCTION

Video analysis covers important research topics such as control of private areas, person tracking, crowd analysis, crowd simulation, anomaly detection etc [1]. By the rapid growth of the public video surveillance systems, the automatic processing of crowd behavior became a popular research area. In recent years both automatic and semi-automatic approaches has been developed by researchers. Ali and Shah [2,3] developed a system which determines the entry and exit points of a high density crowd by using moving averages. They also detected the obstacles such as walls and the most active individuals in the crowd. Mao et al. offered a four step tracking free solution for the crowd analysis. They first extracted the motion parts, and then fine-tuned these parts by using low level features. To correct the errors caused by perspective and occlusion they applied a weight assignment and finally estimated the pedestrian number with various regressors [4]. In [5] the authors offered a real time crowd density estimator based on Markov Random Fields. Li et al. [6] used a self-organizing map to determine the crowd density. They first removed the background by using optical flow and then extracted a set of new features. Ge and Collins [7] modeled the crowd with marked point process and then estimated the number of individuals with Bayesian methods. In another efficient method that [8] estimates information of crowd movement by using the optical flow and the numbers of moving objects by using the numbers of edge pixels. The methods classify the crowd density using the two estimated information according to five levels of crowd density. To overcome the deficiencies of the existing methods used in the estimation of the crowd flow with high-density and multi-motion direction, a crowd flow estimation method based on dynamic texture and generalized regression neural network (GRNN) is presented in the Yu et. al papers [9]. The method firstly extracts the dynamic texture features through optical flow, performs the moving crowd segmentation by the dynamic texture features and level set algorithm to achieve ROIs, and then the regression analysis based on GRNN between ROI features and crowd flow is adopted to achieve the real-time crowd flow estimation results in the crowd scene. Hsu, Lin and Tsai [10] adopt a low cost camera to gather visual data and propose a cellular model for data interpretation. Based on the model, the motion status of the measured area can be represented as a dynamic state matrix, so the proposed method can save a lot of computing time. They adopted the Discrete Cosine Transformation to transform the motion status of the measured area into the frequency domain to recognize the frequency distribution. After then, the feature values are extracted based on different frequency bands and distinct directional information to form a feature vector for training and classification. Finally, the Support Vector Machine is used to classify the feature vector into five classes of crowd density, with the results showing in their system is highly effective in crowd monitoring. Yang and Zhao [11] proposes a new approach used for crowd density estimation. First, background is removed by using a combination of optical flow and background methods. Then according to texture analysis, a set of new features are extracted from foreground image. Finally, a self-organizing map neural network is used for classifying different crowds.

Subburaman et. al [12] detect the head region since this is the most visible part of the body in a crowded scene. The head detector is based on state-of-art cascade of boosted integral features. To prune the search region they propose a novel interest point detector based on gradient orientation feature to locate regions similar to the top of head region from gray level images. Two different background subtraction methods are evaluated to further reduce the search region. Chan and Vasconcelos's approach to the problem of estimating the size of inhomogeneous crowds, which are composed of pedestrians that walk in different directions, without using explicit object segmentation or tracking is proposed. Instead, the crowd is segmented into components of homogeneous motion, using the mixture of dynamic-texture motion model. A set of holistic low-level features is extracted from each segmented region, and a function that maps features into estimates of the number of people per segment is learned with Bayesian regression [13].

978-1-5090-5795-5/17/$31.00 ©2017 IEEE

Fradi and Dugelay in 2013, proposed approach consisting of generating fully automatic and crowd density maps using local features as an observation of a probabilistic crowd function. It also involves a feature tracking step which allows excluding feature points belonging to the background [14]. Another approach [15, 16] proposes a novel approach for crowd density estimation.

The main contribution of this paper is two-fold: First, we propose to estimate crowd density at patch level, where the size of each patch varies in such way to compensate the effects of perspective distortions; second, instead of using raw features to represent each patch sample, we propose to learn a discriminant subspace of the high-dimensional Local Binary Pattern (LBP) raw feature vector where samples of different crowd density are optimally separated. In [17] authors according to the gradual change of crowd density and risk probability in daily hours and uncertainty in their knowledge in evaluation of crowded places, they designed a fuzzy decision making system to make decisions about risk probability. The design of this system is based on the fact that the human visual system tends to direct attention to events that happen with low probability. The efficiency of this system is tested on real data and results are presented to demonstrate the practical applications of this system to aid the human operator. Rao et al. [18] presented a block-based dense optical flow with spatial and temporal filtering is used to obtain velocities in order to infer the locations of objects in crowded scenarios. Furthermore, a hierarchical clustering is employed to cluster the objects based on Euclidean distance metric. The Cophenetic correlation coefficient for the clusters highlighted the fact that our preprocessing and localizing of object movements form hierarchical clusters that are structured well with reasonable accuracy without temporal post-processing. In another approach from Ping et al. [19] combine based on pixel statistics, and the other is based on texture analysis methods and cut the image pixels into small blocks, for low density blocks they adopt the rapid pixel statistics, and for high density blocks they use the combination of pixel statistics and texture analysis to get better estimation. In 2014 Yuan proposed a crowd monitoring approach using mobile phone. their design of crowd detection adopts clustering methods. Feature sets derive from Wi-Fi signal strength measurements. use Bluetooth readings analyzing to estimate crowd density. They implement our design on off-the- shelf smartphones and evaluate its performance via extensive experiments in typical realworld scenes [20]. Karpagavalli and Ramprasad [21] proposed method divided into two fold. In first, we propose density estimation of the crowd size. Secondly, count the number of people in the crowd. As crowd density increases, the occlusion between the people also increases. In order to avoid such problem in crowd we can use Improved Adaptive K-GMM Background subtraction method to extract the exact foreground in real time applications to avoid the estimation problem. By applying boundary detection algorithm, they can estimate the size of the crowd. The number of people in a crowd is counted by using algorithm “canny edge detector”, “connected component labeling” method and “bounding box with centroid” method. They proposed a real time video surveillance system. Wu et al. [22] presented the crowd density estimation for regions based on regional feature analysis and support vector regression (SVR). They extract the following features from each segmented region: the pixel ratio and block-size histogram of the foreground, the pixel ratio and the Minkowski dimension of the edge image, and the gray-level co-occurrence matrix (GLCM) features of the gray image. SVR is used to train and estimate the crowd densities of regions with the extracted features. Secondly, they use the estimated crowd densities and the Lucas-Kanade (LK) optical flow to count pedestrians passing through the line. Then, people moving speeds based on the LK optical flow are used to compute the number of foreground pixels crossing the line.

II. METHODS

In this work, a combined crowd density estimation method is presented. This section describes the elements of the proposed method briefly.

A. SIFT

SIFT is a powerful tool for connected component analysis. By using SIFT it is possible to analyze, track and measure both the moving and static objects in a video. The first step of SIFT is to construct the blurred images by convolving the current image with Gaussian kernels:

$$L(x, y, k\sigma) = G(x, y, k\sigma) * I(x, y), \quad (1)$$

where G(.) is a Gaussian kernel function such as

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

Once the blurred images are found, the differences of Gaussians (DoG) are calculated.

$$DoG = D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (3)$$

The local minima and maxima are called as key points. If the second order Taylor expansion of location $\hat{x}$ is less than 0.3, this point is discarded. To increase the stability, the points which are poorly determined but have high edge responses are eliminated by using Hessian matrix H. For a given threshold r_{th}, if $Tr(H)/\det(H) > (r_{th}+1)^2/r_{th}$ then this key point is rejected. Once the key points are calculated the orientations of key points are calculated:

$$m(x, y) = \sqrt{[L(x+1, y) - L(x-1, y)]^2 + [L(x, y+1) - L(x, y-1)]^2} \quad (4)$$

$$\theta(x, y) = tan^{-1} \frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)} \quad (5)$$

For each 16x16 regions, 4x4 orientation histograms are calculated by using 8 neighbors. Hence, we have 4x4x8 = 128 features [23].

B. Wavelet transform

The optical flow problems require a precision lesser than one pixel. To overcome this situation Magarey and Kingsbury [24] offered a method called complex discrete Wavelet transform. Unlike the SIFT, the discrete wavelet transform (DWT) is not robust to shift and rotation. However, the complex wavelet transform (CWT) overcomes these issues. Basically CWT is a complex valued extension of DWT. 2-D CWT is achieved by separable filtering along first rows then columns (Fig.1).

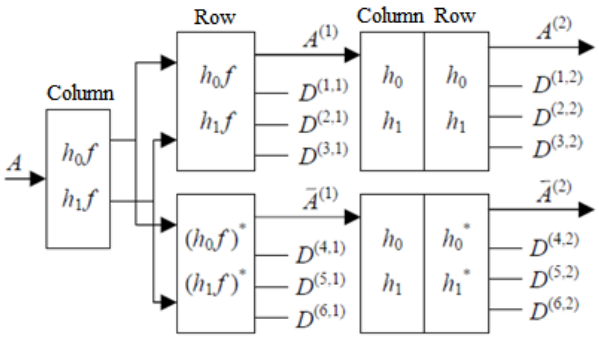

Fig 1. 2-Dimensional Complex Discrete Wavelet Transform [25].

1-D filters h_0 and h_1 are Gabor filters which have complex valued kernels:

$$h_0(n) \approx a_0 e^{-\frac{1}{2}\left(\frac{n+0.5}{\sigma_0}\right)^2} e^{j\omega_0(n+0.5)}, \quad \text{and}$$

$$h_1(n) \approx j a_1 e^{-\frac{1}{2}\left(\frac{n+0.5}{\sigma_1}\right)^2} e^{j\omega_1(n+0.5)} \quad (6)$$

For the lower frequency rows the complex conjugates of these filters are used. Hence for each level, six complex band-pass signals and two low-pass signals are obtained. The movement is estimated with subband squared difference:

$$SSD^{(j)}(n,f) = \kappa \sum_{s=1}^{6} \left| D_1^{(s,j)}(n+f) - D_2^{(s,j)}(n) \right|^2 \quad (7)$$

where D_i are details of i^{th} frame, f is a real valued offset, j is the level of wavelet transform and s is the index of the subband which is currently considered. The algorithm is applied to all sub bands starting from the lower band. The cumulative of the calculated SSDs is the motion surface [24, 25].

C. *Principal component analysis*

Eigenvectors of a given matrix define an orthonormal vector space. In the principal component analysis (PCA), it is assumed that the eigenvector which has the highest eigenvalue is more distinctive than the others. Similarly, the eigenvector which has the least eigenvalue is the least distinctive, or redundant. If these redundant vectors are omitted, the remaining are considered as distinctive features.

D. *Kalman filter*

Kalman filter is an adaptive linear estimator that works on a state space. According to Kalman filter, the true state at time n is calculated from the state at (n-1):

$$x(n) = \Phi . x(n-1) + \Gamma u(n) + w(n), \quad (8)$$

where x(n) is the state variable, Φ is the state transition model, u(n) is the filter coefficients, w(n) is the process noise and Γ is the control input model applied to u(n). At any time, the measured value z(n) of the real state x(n) is

$$z(n) = H . x(n) + v(n), \quad (9)$$

where H(n) is the measurement matrix and v(n) is the measurement noise. v(n) is considered as Gaussian with zero mean. Fig 2 shows the general model of Kalman filter [26]. In the object tracking problems, any preprocessing which estimates the starting positions and sizes of the objects being tracked dramatically improves the performance of Kalman filter.

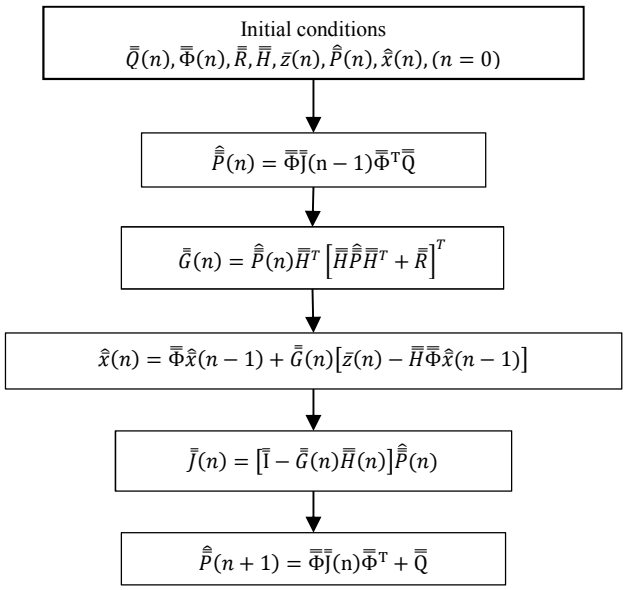

Fig 2. Kalman Filtering Algorithm.

III. APPLICATION

In this work, three common crowd analysis benchmark data which PETS2009 [28], UCSD [29] and Grand Central [30] are considered. PETS2009 data consist of shots taken from a campus with eight cameras. The data divided into four parts labelled as S0, S1, S2 and S3. S0 is the training data and it contains three sets of video sequences including background models of each cameras, random walking crowd flow and regular walking crowd. S1 set is used for person count and density estimation. S2 set is used for people tracking. S3 is suitable for flow and event analysis. In this work, we used Time13-57 and Time 13-59 parts of S1 data. UCSD data is taken by using a static 8-bit gray scale camera in 10 fps. The data split into two parts taken from two different viewpoints. In our experiments the first 2000 annotated frames of UCSD data are considered. Grand Central dataset is recorded at New York Grand Central Train Station. The whole length of the video is 33 mins with 25 fps speed. We used all of this dataset. Figure 3 shows sample frames from these datasets. Table I summarizes the datasets.

TABLE I. OVERVIEW THE CONDITIONS OF THE THREE DATASETS

	PETS2009	**USCD**	**Grand Central**
Frames	1565	2000	46009
Frame Rate (Fps)	7	10	23
Resolution	768x579	236x158	720x480
Color	RGB	Gray	Gray
Location	Outdoor	Outdoor	Indoor
Shadow	Yes	No	No
Reflexion	No	No	Yes
Loitering	No	No	Yes
Crowd Size	0-42	13-53	125-245

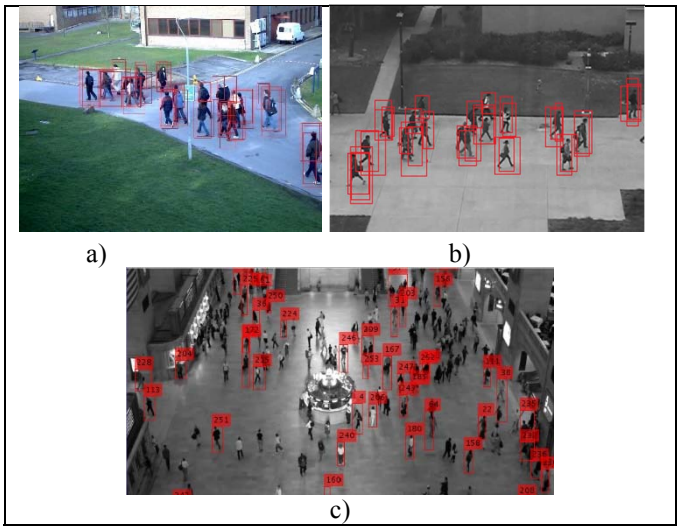

Fig 3. Sample frames from the datasets. a) PETS2009, b)UCSD, c) Grand Central.

The general flow of the proposed method is given in Figure 3. In the first step, a blob analysis based on SIFT is applied to frames. The poorly determined points are eliminated by using DoG. A CWT is applied to these blobs to obtain the movement directions and the speeds of key points. Similarly, the speed and directions of the blobs are calculated by using optical flow. To reduce the computational complexity, a PCA is applied to blobs before the calculation of optical flow. If any blob appears both in the result of CWT and optical flow, it is considered as a person. One of the main problems in people tracking/counting application is occlusion. To track the occluded people's trajectories Kalman filtering is applied to final people map. If an occluded person is not being tracked again in an acceptable time period, the person is assumed to be left the scene.

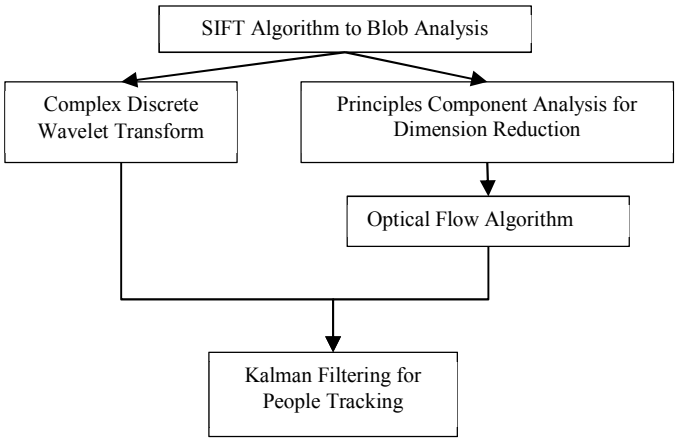

Fig 4. Proposed Algorithm.

The mean absolute error (MAE) term is used as the performance metric to determine the performance of the system

$$MAE = \frac{1}{T}\sum_{i=1}^{T}|n_i - \hat{n}_i|, \qquad (20)$$

where T is the total amount of test frames, n_i is the ground truth of the frame i and $\hat{n}_i$ is the estimated crowd count of the frame i. Table II shows our results regarding to recent works based on the same data. As seen in the Table II, our proposed method produced better results for PETS2009 and Grand Central data. For the UCSD data, our method is among the best three.

TABLE II. PERFORMANCE COMPARISON BETWEEN DIFFERENT METHODS AND PROPOSED METHOD

	Mean Absolute Error		
	PETS 2009	***UCSD***	***Grand Central***
Ma et al. [31]	2,55	3,60	9,57
Lempitsky and Zisserman [32]	2,62	2,60	8,33
Chen et al. [33]	2,78	2,23	8,40
Xu at el. [34]	2,54	2,23	8,12
Zhang et al. [35]	-	1,60	-
Topkaya et al. [36]	1,47	1,30	-
Proposed Method	1,37	1,63	7,98

IV. CONCLUSION

In this work a ccombined crowd density estimator is presented and the performance of the presented method tested on three well known data. MAE term was used as performance metric. By comparing the literature, the proposed method gave better MAE for PETS2009 as 1.37 and Grand Central data as 7.98. For the UCSD data the best MAE is reported in [36] as 1.30 whereas our method's score is 1.63. By considering the structural differences in the videos, one may claim that the proposed method has better generalization ability.

REFERENCES

[1] Jacques Junior, J.C. S., Musse, S. R. and Jung, C. R., "Crowd Analysis Using Computer Vision Tecniques," IEEE Signal Processing Magazine, Vol. 27, pp. 66-77, 2010.

[2] Ali, S., Shah, M., "A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis," in IEEE Conference on Computer Vision and Pattern Recognition, pp. 2054-2060, 2010.

[3] Ali, S., Shah, M., "Floor Fields for Tracking in High Density Crowd Scenes," 10th European Conference on Computer Vision-ECCV, Lecture Notes in Computer Science Volume 5303, 2008, pp 1-14, 2008.

[4] Mao, Y., Tong, J. and Xiang, W., "Estimation of Crowd Density Using Multi-Local Features and Regression," Proceedings of the 8th World Congress on Intelligent Control and Automotion, pp. 6295-6300, 2010.

[5] Guo, J., Wu, X., Cao, T., Yu, S., and Xu, Y., "Crowd Density Estimation via Markov Random Field (MRF)," Proceedings of 8th World Congress on Intelligent Control and Automation, pp. 258-263, 2010.

[6] Li, W., Wu, X., Matsumoto, K. and Zhao, H., "A New Approach of Crowd Density Estimation," IEEE Region 10 Conference TENCON, pp. 200-203, 2010.

[7] Ge, W. and Collins, R. T., "Crowd Density Analysis with Marked Point Processes," IEEE Signal Processing Magazine, Vol. 27, pp. 107-123, 2010.

[8] Kim, G., Eom, K., Kim, M. and Jung, J., "Automated Measurement of Crowd Density Based on Edge Detection and Optical Flow," IEEE 2nd International Conference on Industrial Mechatronics and Automation, Vol. 2, pp. 553-556, 2010.

[9] Yu, H., He, Z., Liu, Y. and Zhang, L., "A Crowd Flow Estimation Method Based on Dynamic Texture and GRNN," 7th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 79-84, 2012.

[10] Hsu, W., Lin, K. and Tsai, C., "Crowd Density Estimation Based on Frequency Analysis," 7th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 348-351, 2011.

[11] Yang, H. and Zhao, H., “A Novel Method for Crowd Density Estimations,” IET International Conference on Information Science and Control Engineering (ICISCE), pp. 1-4, 2012.

[12] Subburaman, V. B., Descamps, A. and Carincotte, C., “Counting People in the Crowd Using a Generic Head Detector,” IEEE 9th International Conference on Advenced Video and Signal-Based Surveillance, pp.470-475, 2012.

[13] Chan, An. and Vasconcelos, N., “Counting People with Low-Level Features and Bayesion Regression,” IEEE Transactions on Image Processing, Vol. 21, No. 4, pp. 2160-2177, 2012.

[14] Fradi, H. and Dugelay, J., “People Counting System in Crowded Scenes Based on Feature Regression,” Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 136-140, 2012.

[15] Fradi, H., Dugelay, J., “Crowd Density Map Estimation Based on Features Tracks,” IEEE 15th International Workshop on Multimedia Signal Processing, pp. 40-45, 2013.

[16] Fradi, H., Zhao, X. and Dugelay, J., “Crowd Density Analysis Using Subspace Learning on Local Binary Pattern,” IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1-6, 2013.

[17] Tehranipour, F., Shishegar, R., Tehrenipour, S. and Seterehdan, S., “Attention Control Using Fuzzy Inference System in Monitoring CCTV Based on Crowd Density Estimation,” IEEE 8th Iranian Conference on Machine Vision and Image Processing (MVIP), pp. 204-209, 2013.

[18] Rao, A. S., Gubbi, J., Marusic, S., Stanley, P. and Palaniswami, M., “Crowd Density Estimation Based on Optical Flow and Hierarchical Clustering,” IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 494-499, 2013.

[19] Ping, K., Bo, P., Wenying, Z. and Shuai, L., “Research on Central Issues of Crowd Density Estimation,” 10th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 143-145, 2013.

[20] Yuan, Y., “Crowd Monitoring Using Mobile Phones,” IEEE 6th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Vol. 1, pp. 261-264, 2014.

[21] Karpagavalli, P. and Ramprasad, A. V., “Estimating the Density of the People and Counting the Number of People in a Crowd Environment for Human Safety,” IEEE International Conference on Communication and Signal Processing, pp. 663-667, 2013.

[22] Wu, Z., Zheng, H. and Wang, J., “Pedestrian Counting Based on Crowd Density Estimation and Lucas-Kanade Optical Flow,” IEEE 7th International Conference on Image and Graphics (ICIG), pp. 471-476, 2013.

[23] Lowe, David G., "Object Recognition From Local Scale-Invariant Features," Proceedings of the International Conference on Computer Vision. pp. 1150–1157, 1999.

[24] Megarey, J. and Kingsbury, N., “ Motion Estimation Using a Complex-Valued Wavelet Transform,” IEEE Transactions on Signal Processing vol. 46, no. 4, pp. 1069-1084, 1998.

[25] Yılmaz, Ş. and Severcan, M., "Target Tracking Using The Complex Discrete Wavelet Transform Based Motion Estimation Method," Proceedings of the IEEE Signal Processing and Communications Applications Conference, pp. 605-608, 16-18 May 2005.

[26] Haykin S., “Adaptive Filter Theory”, Prentice Hall Information and System Science Series,1996.

[27] Temiz, M. S., “Video Görüntülerinden Hareketli Nesnelerin Çıkarılması ve Hareket Yörüngelerinin Belirlenmesi”, Doktora Tezi, İstanbul Teknik Üniversitesi, Jeodezi ve Fotogrametri Anabilim Dalı, Geomatik Mühendisliği, Aralık 2011.

[28] Antoni B Chan, Mulloy Morrow, and Nuno Vasconcelos, “Analysis of crowded scenes using holistic properties,” in Performance Evaluation of Tracking and Surveillance workshop at CVPR, 2009, pp. 101–108.

[29] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos,“Privacy preserving crowd monitoring: Counting people without people models or tracking,” in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008, pp. 1–7.

[30] Bolei Z., Xiaogang W., and Xiaoou T., “Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrianagents,” in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 2871–2878.

[31] Ma W., Huang L., and Liu C., “Crowd density analysis using co-occurrence texture features,” in Computer Sciences and Convergence Information Technology (ICCIT), 2010 5th International Conference on. IEEE, 2010, pp. 170–175.

[32] Lempitsky V. and Zisserman A., “Learning to count objects in images,” in Advances in Neural Information Processing Systems, 2010, pp. 1324–1332.

[33] Chen K., Loy C. C., Gong S., and Xiang T., “Feature mining for localised crowd counting.,” in BMVC, 2012, vol. 1, p. 3.

[34] Xu T., Chen X., Wei G., and Wang W., “Crowd Counting Using Accumulated HOG”, 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016.

[35] Zhang C., Li H., Wang X., Yang X., “Cross-scene Crowd Counting via Deep Convolutional Neural Networks”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[36] Topkaya İ. S., Erdogan H., Porikli F., “Counting people by clustering person detector outputs”, 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2014.