2023 International Conference on the Confluence of Advancements in Robotics, Vision and Interdisciplinary Technology Management (IC-RVITM) | 979-8-3503-0608-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/IC-RVITM60032.2023.10435190

Detection and Tracking of Multiple Faces in Video using Modified KLT Algorithm

Mr. Dileep J
Research Scholar
Dept. of ECE
Sir MVIT, VTU
Bengaluru, India
dileep1721991@gmail.com

Dr. Supriya V G
Professor & Head
Dept. of ECE
Sir. MVIT
Bengaluru, India
hod_ece@sirmvit.edu

Dr. Manjunath Ramachandra
Professor
Dept. of ECE
Atria Institute of Technology,
Bengaluru, India
drmanjunathramachandra@gmail.com

Abstract*— In proposed research, A system for accurate face detection and continual monitoring that makes use of face key characteristics as essential elements of the face that are incorporated in modified KLT approach. A greater degree of data is provided by video frames than by a single image. Critical attribute extractors (points) are used to automate the detection and monitoring of faces in a wide range of applications, such as human behavior identification, automated surveillance and human computer interface. By dividing the monitoring approach into three sections, one can construct a straightforward facial detection and continual monitoring system. Face detection, facial feature identification and tracking of faces are the first three steps. The Viola Jones object detector is used to achieve the face detection and modified KLT method is used to accomplish the continual monitoring. According to the findings of the investigation, our method is more reliable and effective at evaluating video frames to account for different types of occlusions and pose variations. Visual Tracker Benchmark database is used as a standard database. Further results are compared with staff database which was created to verify the outcomes. Average accuracy of 95% is obtained for the proposed method.

Keywords — Alignment Estimation, Detection, Modified KLT Algorithm, Tracking

I. INTRODUCTION

Since quite a while ago, the investigators have experienced a significant barrier in identifying and recognising the presence of faces as well as other objects. It is a challenging issue in artificial intelligence techniques and a significant field of study. Utilisation possibilities like biometric authentication, sophisticated human computer interface, surveillance, and interactions between humans and robots all heavily rely on face detection[1]. The ability to categorise typical and unusual events in real time using facial features for recognition and monitoring is becoming more and more advanced in applications connected to safety and confidentiality. Signals generated in robots or machines can be successfully directed and indicated by these kinds of monitoring tools. Due to numerous variables and ever-changing designs, such as angle of viewpoint, position differences occlusions, face viewpoints, backdrop intensity, and illumination parameters, face recognition and monitoring in video poses a significant research challenge[2].

In order to address the issue of face recognition and monitoring in films, multiple scientists have developed many methods, including 1) Knowledge-based methods, 2) Feature invariant strategies, 3) template-matching approaches and 4) Appearance-based methods. Each technique has benefits and disadvantages. Any system for identifying and monitoring faces will occasionally show aberrations that stray from the desired object of investigation. Any tracking technique can only be effective and reliable if it can reduce the dispersion from the desired result and handle the difficulties in interpreting the videos. In this paper, optimised object designs and the Viola-Jones approach for detection was used. We use the Kanade-Lucas-Tomasi (KLT)[3] algorithm to monitor a person's face from frame to frame in videos since it is more efficient and accurate than many other methods currently in use.

II. RELATED WORK

Trackers frequently fail to make reliable identifications in environments that are crowded. In [1], author suggested a novel MOT methodology for cluttered environments called tracking-by-counting. author concurrently described several target identifications, counting, and monitoring as a distributed flow programme using crowd density maps. This programme simultaneously determined universal optimised findings and itineraries of multiple targets across the entire video. This is in contrast to previous MOT approaches, which either rely on a poor two-step procedure using intuitive density aware point-tracks for pairing destinations or neglect the crowd density and are hence susceptible to glitches in crowded settings. On public acceptable standards for a variety of domains, such as people monitoring, cell monitoring, and fish tracking, this method produced encouraging results. Due to extended (short) obstruction between targets, ID-switch occurs when many targets are tracked via identification in video clips. To address this problem, author in [2] suggested a Recurrent Neural Network (RNN)-based obstruction managing approach in this research. By anticipating the findings in subsequent frames, the approach rebuilds failed boxes so as to maintain the identification numbers of entities following obstruction. By employing a cutting-edge RNN to acquire target activity, an estimate is made. The monitoring findings of numerous cutting-edge technologies using this process demonstrate that their ID-switch mistake is decreased.

To reduce distortion and improve face recognition using a distance-based classification algorithm, facial characteristics are identified and recorded using the 2DPCA attribute extraction methodology [3]. The Face94 database and webcam photos are utilised to test the proposed method. An interpretation approach integrating low-level and mid-level characteristics is suggested in [4], which recognises a gesture from the coarse to the fine design, in an effort to address the issue of the low identification rate of an individual low-level characteristic for multi-view facial movement recognition. system first obtains mid-level characteristic depended LLC (locality-constrained linear coding) on facial regions that are active using conventional SPM. Then, on the entire face, author estimated the PHOG Describer as a low-level

characteristic. The low-level and mid-level characteristics are then concatenated, which is a straightforward. Author tested method thoroughly using the SDUMFE and Multi-PIE records, and the findings reveal that it provides excellent outcomes for multi-view facial movement detection. In [5], author suggested an integrated resilient tumble extrapolation architecture that is capable of handling both pictures with extreme obstruction and photos with oversized head postures. The approach precisely forecasts the landmark positions and landmark obstructions using iterative prediction. In order to enhance the effectiveness of blockage prediction, author further explicitly specify occluded pattern as a restriction. In order to iteratively update the positions of the landmarks, system mixed the probability of landmark accessibility, local perceptions, and local shapes.

Target monitoring effectiveness is frequently negatively impacted by the complicated surroundings. To reduce the disruptive effects brought on by surroundings monitoring, the motion of the target's features can be removed. In order to choose the area to be tracked, also known as the space of concern, and to separate it from the backdrop, a gauss weight operation is used [6]. The backward estimate plan of the area to be studied is subsequently monitored separately, removing the background's potential interference with the object. The empirical findings demonstrate that the suggested Camshift procedure that utilizes multi combined features has greater precision and dependability contrasted to the obsolete Camshift monitoring procedure. The suggested approach is contrasted with the obsolete Camshift procedure and the Camshift technique dependent on multi feature integration. A unique Camshift monitoring strategy that utilizes edge diminution has been suggested to enhance the target's monitoring effectiveness against intricate backgrounds [7]. Given that the target's edge is likely to be most susceptible to background interference, the weighted approach is used to minimize the target edge's grey value in the back estimation picture based on its spot and dimension in the preceding frame. As a result, it is possible to stop ambient disruption knowledge from interfering with target monitoring. The technique is used to follow the target in a recording of the common test libraries as well as the actual target. According to experimental findings, suppressing the edge information of the tracking target in the rear projection image can increase the precision and consistency of target identification and monitoring.

In [8], author described a deep regression method for symmetry of faces. A neural system called the deep regressor has a larger section and many local regions. While the subsequent regional stages progressively refine the form using closer picture findings, the global layer predicts the starting shape of the face from the entire picture. By preventing a pattern that frequently occurs in cascaded predictive techniques and degrades performance as a whole, author demonstrated that the outcomes as deep regressor progressively and uniformly arrived at the real features of the face stage by stage. This avoids the tendency that produces initial phase regression models that boast substantial position precision gains but thereafter regressors with low conformity precision profits. The article [9] described a technique for cluster monitoring using lasers. Images are produced via a two-layer scanning laser that is integrated into the ecosystem. The photos' human position information is extracted using a backdrop subtraction technique. In congested surroundings, information relationship follows pragmatic criteria and the global nearest neighbor (GNN) recognizes multiple individuals. In [10], author offered a method for monitoring numerous persons online using a single camera. Support vector machine (SVM) specific to individuals' classification techniques, resemblance ratings, the Hungarian technique, and inter-object obstruction handling had all been combined in the monitoring phase. The suggested method has no prerequisites and places no restrictions on the surrounding environment.

In [11], a reliable approach for monitoring persons in videos taken by several cameras is proposed. Author provided an approach for improving resilience through the use of various attribute kinds. Even without these functionalities, author can still follow individuals. According to the research results, an average of 76% of the pictures taken from the identical subject were recognized. Author in [12] offered an innovative solution to the issue of simultaneously monitoring and recognizing several individual's gaits in a clip of video. The sub-problems of multiple individual monitoring and video gait identification can both benefit from each other. Therefore, as an architecture to increase gait detection precision and reduce ID switching in monitoring, we suggest joint monitoring and gait identification for several individuals. An effective surveillance system that monitors plenty of individuals concurrently and intelligently is shown in [13]. To monitor the people who seem to be intruders, a unified framework of Complex Event Processing (CEP) and Kalman monitoring is adopted. In case, an individual is detected by sensors, statistics or instances are established. There was a lot of disturbances, replication, and incomplete details within the data produced in a WSN (Wireless Sensor Network) context. The suggested solution uses CEP, a knowledge base, and Kalman monitoring to take care of each of these problems.

Author in [14] described that, a person's tracking system that combines a fisheye camera and a two-layered laser range sensor (LRS). Individual's waists and knees are recognized by the LRS, and the camera records colored photos of the torsos of those individuals. A conversing multidimensional paradigm predictor is used to monitor individual's activities including running, walking, abrupt starting/stopping, and rapid rotation. The classified individuals are monitored using a model-based tracker. With just one camera, author suggest using a Binary/Appearance Tracking System that combines silhouette resemblance subtraction of the background, and fragment filter to estimate pedestrians' whereabouts in a variety of obstruction scenarios [15]. During the obstruction time, binary and colored silhouettes are flexibly employed to determine the similarities between the observation with potential silhouette configurations are. As a result, the precise positions of the obscured pedestrians can be easily determined using every possible conceivable arrangement of silhouettes. In [16], author addressed the topic of monitoring moving individuals with a group of passing robots. The calculation of hypothesis the likelihood takes into consideration both the potential for error in recognition and

the adjoining domains of vision of the robot's sensors. The effectiveness of our monitoring technique's application to two tests comprising individuals transitioning around rolling robots is evaluated. In results and discussion chapter, result comparison table with research gaps is discussed.

III. Proposed Methodology

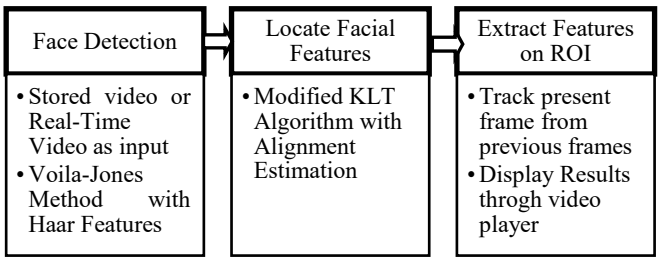

Fig. 1: Proposed Methodology

Fig. 1 shows proposed methodology for simple face detection and tracking method by categorizing the problem in to three parts, namely Face detection, locating facial features and track the present frame from the previous frames. During the face detection phase, the system uses the Cascade object analyzer to determine the face in a video frame to recognise the item of concern. The Viola-Jones technique of identification and an optimised model for classification are both used by the object analyzer. The Viola-Jones approach is used to determine where human faces are located inside video frames. Any face or interesting feature can be found using the analyzer in groups of video frames.

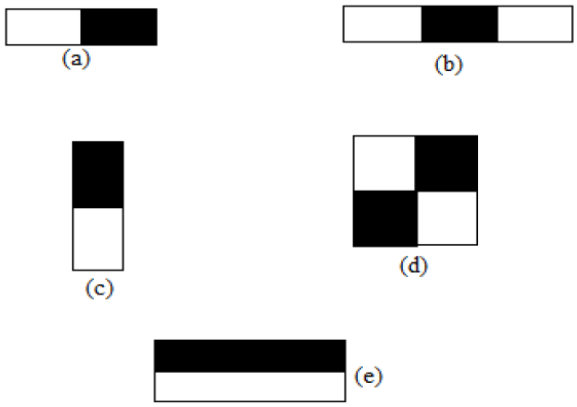

Fig. 2: Haar Features

Haar features are used initially to train the system as an integral part of voila jones framework. Black and white rectangular spaces make up the Haar characteristic features. Haar characteristics are straightforward rectilinear attributes that are the sum of pixels from different places within the defined area. The rectangular shape has the ability to resize the picture and may be placed anywhere in the frame. 2-rectangle characteristic is the name of this revised set of attributes. Every characteristic group can reveal the presence or shortage of specific details like borders or appearance modifications within the frame. For assessing the face attributes, these haar characteristics are used. For instance, in Fig. 2(b), the black-colored portion indicates the existence of a nose, that is situated in the middle of the person's face. This region is utilized to identify this characteristic. By subtracting the entire number of pixels over the white border from the total of pixels over the black rectangle, the outcome is determined. For certain characteristics, a threshold is typically set. Estimated is the average total for every black and white. Next, a threshold verification is done on the disparity in value. The feature is recognized as meaningful if its magnitude exceeds or meets the threshold. System utilizes the Kanade-Lucas-Tomasi (KLT) method to monitor the human face characteristic frame by frame. Since object analyzer are unable to manage difficulties such object displacement and tilting the head by certain angle, they are costly in terms of computing. As a result, it is solely employed to recognize faces, and system uses the KLT method to monitor a group of interesting characteristic points among video clips. After the person's face has been found, the machine learning algorithm ought to recognize the characteristics of the points that have an adequate amount of texture and can be reliably monitored. In our study, the technique for identifying monitored characteristics minimizes the SSD disparity among windows centered at the monitored characteristic point region. This has an appearance that is sufficient to perform automatic monitoring. In order to identify a face, the essential traits are recovered. To figure out the distinguishing characteristics of the human face, the initial frame is then mapped from the previous one using Vision.PointTracker(). For estimating the angle of revolution, modification, and range among the current characteristic points and prior feature points, the Geometric Transform () procedure was constructed. The bounding box surrounding a person's face in video footage also uses it. proposed system creates a point tracker in the next stage and enabling the bilateral loss barrier to strengthen it in a context of distortion and interference. Later, a video player entity will also be created for exhibiting video frames.

Assume that the edge's initial boundary position is (i, j). If it is displaced by a variable vector (t1, t2,...., tn) in the following frame, the resultant edge position in the frame will then be the result of the combination of the initial point and transferred vector. The newly created point's dimensions will be i'=i+t1 and j'=j+t2. Therefore, it is now necessary to determine the displacement in relation to every position. Warp function, an algorithm with positions and an argument, is used for this. The symbol for it is W(i;d) = (i+t1;i+t2). The evolution is estimated using the warp function. The initial identified points are used as an outline picture in the very first frame. In later stages, the monitoring positions are determined by subtracting the displacement from the point before.

The alignment is computed by

$$\sum_i [I(W(i;d)) - T(i)]^2 \qquad (1)$$

Where d is displacement constraint. Considering d as initial estimate which is a known parameter and determining Δd.

$$\sum_i [I(W(i;d+\Delta d)) - T(i)]^2 \qquad (2)$$

Displacement Δd can be determined by calculating Taylor series and taking differential vector with respect to Δd.

$$\Delta d = H^{-1} \sum_{i} \nabla I \left[\frac{\partial W}{\partial d}\right]^{T} . \left[T(i) - I\left(W(i;d)\right)\right] \qquad (3)$$

Where H is called as hessian matrix. This is how we predict the displacement Δd and determine the next traceable point. At last, video player is used to display the resultant video stream.

IV. Results and Discussion

This process is executed using MATLAB tool. A stored video templates can be given as input in any format. (MP4, AVI etc.) Through webcam, live video can be taken as real time input. Visual Tracking Benchmark (VTB) Database had been used as a standard database. In this multiple video files had been taken for analysing proposed framework. Further, Staff database (own) had been created for verifying the work. In that, multiple video files had been taken in 30 frames per second frame rate and 8 Mega Pixel resolution. Table I shows the performance metric considering VTB database. 98.38% of maximum accuracy was obtained considering different challenges. Total number of 62 features had been identified out of which 61 attributes were with respect to face. Fig. 3 and Fig. 4 shows VTB Database results which clearly identifies given number of people with maximum number of attributes. Both coloured and black-white databases are used for testing proposed methodology.

TABLE I. Accuracy of Proposed Method for VTB Database

Videos	*Total no of features identified*	*Total number of non facial features*	*Accuracy (%)*
1	85	10	88.23
2	225	15	93.33
3	62	1	98.38

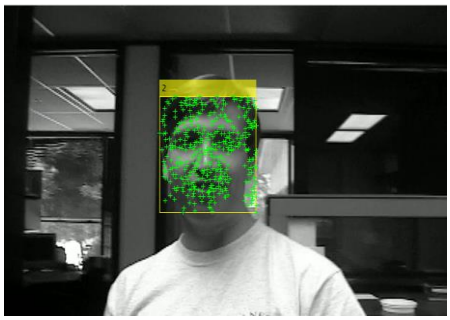

Fig. 3: Visual Tracking Benchmark Database result-1

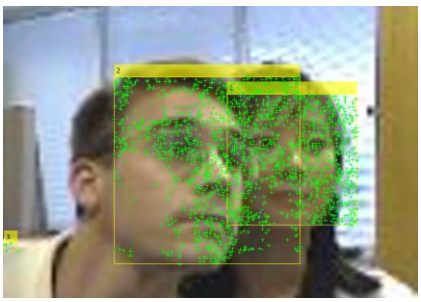

Fig. 4: Visual Tracking Benchmark Database result-2

Table II shows accuracy of proposed methodology by taking staff database into consideration. Maximum precision of 95.55% was achieved while recognising 129 facial attributes out of 135 detected features. 95% of average recognition rate was obtained for face detection and monitoring in wide variety of video frames. Fig. 5 shows the snapshot of staff database result for proposed methodology. Table III shows the performance metrics for real time video obtained through web camera. 98.41% of maximum efficiency was obtained for which it gave only one non-facial attribute. Multiple runs with different lightening conditions were verified. Maximum number of 16 faces had been detected at one shot using proposed methodology, which is highest so far. Fig. 6 and Fig. 7 shows few sample pictures while detecting faces in real time through web camera. It was clear that, all faces had been identified accurately obtaining highest results.

TABLE II. Accuracy of Proposed Method for Staff Database

Videos	*Total no of features identified*	*Total number of non facial features*	*Accuracy (%)*
1	135	6	95.55
2	84	5	94.04
3	68	4	94.11

Fig. 5: Staff Database result

TABLE III. Accuracy measurement for Real Time Video

Videos	*Total no of features identified*	*Total number of non facial features*	*Accuracy (%)*
1	48	3	93.75
2	63	1	98.41
3	110	6	94.54
4	456	35	92.32
5	82	3	96.34
6	128	21	83.59
7	78	10	87.17
8	91	4	95.60

Fig. 6: Real-Time output-1 from web camera

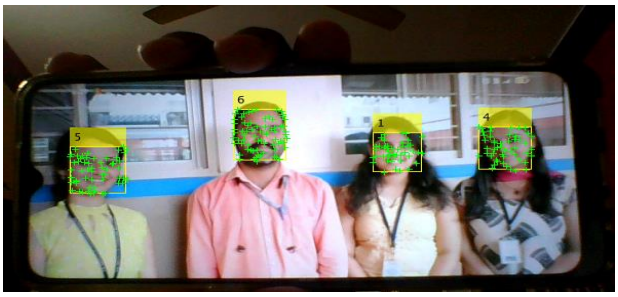

Fig 7: Real-Time output-2 from web camera

TABLE IV. RESULT COMPARISON

Paper	*Methodology Used*	*Accuracy (%)*	*Research gap*
[1]	Density Maps: Deep learning model	90	Execution time is more. (Minimum of 16 seconds)
[2]	Recurrent Neural Network	94	Pose variation challenge
[8]	cascaded regression algorithm with backpropagation	92.5	Can not identify faces in noisy frame
[10]	Support Vector Machine	94.2	Limited range of detection
[12]	Gait Hypotheses and Support Vector Machine	96	Angle of Pose capture
[14]	heuristic-rule-based and global-nearest-neighbor-based data association and LASER range sensor	90	Model is expensive and slow in identifying moving objects
[16]	Multiple Hypothesis Tracking method	75	Complex structure and occlusion problem
Proposed Method	**Modified KLT Approach with alignment estimation**	**98.41**	**---**

Table IV shows the result comparison and research gap identification in the field of multiple face identification and monitoring. It was evident that, in every paper one or the other limitation with respect to human face detection and monitoring. Proposed method outperforms other existing methods by obtaining maximum accuracy of 98.41% in robust conditions and stated challenges like angle of face capture, different lighting conditions and obstruction.

CONCLUSION

When evaluated against existing approaches, the technique suggested for automated face identification and monitoring in our paper minimizes the predicted calculation time while still producing excellent accuracy. The detection of faces is performed using the Viola-Jones cascaded object detector, and monitoring of faces in a series of video frames is performed using the modified KLT technique while minimizing the efforts in alignment estimation. The majority of surveillance and safety platforms can establish real-time video organizing and categorization requirements, and necessary creatures can be followed effectively. Numerous individuals in one image along with changes in magnitude and position can all be handled using our technique. The findings of investigations conducted on a sizable number of raw data sets clearly outperform those of feature-based identification methodologies. Considering VTB database, staff database and real-time video sequences through web camera, 95% of mean accuracy is maintained.

REFERENCES

[1] Weihong Ren , Xinchao Wang , Jiandong Tian, Yandong Tang and Antoni B. Chan, "Tracking-by-Counting: Using Network Flows on Crowd Density Maps for Tracking Multiple Targets," IEEE Transactions on Image Processing, ISSN: 1941-0042, Vol. 30, pp. 1439-1452, June 2021.

[2] Maryam Babaee, Zimu Li and Gerhard Rigoll, "Occlusion Handling in Tracking Multiple People using RNN," IEEE, International Conference on Image Processing, DOI: 978-1-4799-7061-2, pp. 2715-2719, 2018.

[3] Nawaf Hazim Barnouti, Mohanad Hazim Nsaif Al-Mayyahi, and Sinan Sameer Mahmood Al-Dabbagh, "Real-Time Face Tracking and Recognition System Using Kanade-Lucas-Tomasi and Two-Dimensional Principal Component Analysis," IEEE International Conference on Advanced Science and Engineering, Kurdistan Region, Iraq, DOI: 978-1-5386-6696-8, pp. 24-29, 2018.

[4] Mingyue Bi, Xin Ma, Rui Song, Xuewen Rong, Yibin Li, "Multi-view Facial Expression Recognition Based on Fusing Low-level and Mid-level Features," Proceedings of the 37th Chinese Control Conference, Wuhan, China, pp. 9083-9088, July 2018.

[5] Yue Wu, Qiang Ji, "Robust Facial Landmark Detection under Significant Head Poses and Occlusion," IEEE International Conference on Computer Vision, IEEE Computer Society, ISSN: 1550-5499, DOI: 10.1109/ICCV.2015.417, pp. 3658-3666, 2015

[6] Chunbo Xiu, Fushan Ba, "Target Tracking Based on the Improved Camshift Method," IEEE, 28th Chinese Control and Decision Conference (CCDC), ISSN: 978-1-4673-9714-8, pp. 3600-3604, 2016

[7] Chunbo Xiu, Zuohong Chai, Huiyao He, Jianguo Hou, "Improved CamShift Tracking Method Based on the Edge Suppression," IEEE, 29th Chinese Control And Decision Conference (CCDC), DOI: 978-1-5090-4657-7, pp.3529-3533, 2017.

[8] Baoguang Shi, Xiang Bai, Wenyu Liu, "Face Alignment With Deep Regression," IEEE Transactions on Neural Networks and Learning Systems, ISSN: 2162-237X, pp. 1-12, 2016.

[9] Masafumi Hashimoto, Azusa Nishio, Atsushi Tsuji and Kazuhiko Takahashi, "Laser-Based Tracking of Groups of People with Sudden Changes in Motion," IEEE, DOI: 978-1-4799-7800-7, pp 315-320, 2015.

[10] Sahar Rahmatian, Reza Safabakhsh, "Online multiple people tracking-by-detection in crowded scenes," IEEE, 7th International Symposium on Telecommunications, DOI: 978-1-4799-5359-2, pp. 337-342, 2014.

[11] Satoshi Yoshida, Jianquan Liu and Shoji Nishimura, "A Robust People Tracking Method in Multiple Cameras," IEEE Fifth International Conference on Multimedia Big Data, DOI: 978-1-7281-5527-2, pp. 305-308, 2019.

[12] Maryam Babaee, Gerhard Rigoll and Mohammadreza Babaee, "Joint Tracking and Gait Recognition of Multiple People in Video," IEEE, International Conference on Image Processing, DOI: 978-1-5090-2175-8, pp. 2592-2596, 2017.

[13] R. Bhargavi, K. Sri Ganesh, M. Raja Sekar, P. Rabinder Singh and V. Vaidehi, "An Integrated System of Complex Event Processing and Kalman Filter for Multiple People Tracking in WSN," IEEE International Conference on Recent Trends in Information Technology, MIT, Anna University, Chennai, DOI: 978-1-4577-0590-8, pp. 890-895, 2011.

[14] Masafumi Hashimoto, Zhitao Bai, Tomoki Konda and Kazuhiko Takahashi, "Identification and Tracking Using Laser and Vision of People Maneuvering in Crowded Environments," IEEE, DOI: 978-1-4244-6588-0, pp. 3145-3151, 2010.

[15] Hsin-Ho Yeh, Jiun-Yu Chen, Chun-Rong Huang and Chu-Song Chen, "An Adaptive Approach for Overlapping People Tracking based on Foreground Silhouettes," Proceedings of IEEE 17th International Conference on Image Processing, DOI: 978-1-4244-7994-8, pp. 3489-3492, September 2010.

[16] Nicolas A. Tsokas and Kostas J. Kyriakopoulos, "Multi-Robot Multiple Hypothesis Tracking for Pedestrian Tracking with Detection Uncertainty," 19th IEEE Mediterranean Conference on Control and Automation Aquis Corfu Holiday Palace, Corfu, Greece, DOI: 978-1-4577-0123-8, pp. 315-320, June 2011.