Proceedings of the 10th INDIACom; INDIACom-2016; IEEE Conference ID: 37465
2016 3rd International Conference on "Computing for Sustainable Global Development", 16th - 18th March, 2016
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA)

Efficient Stereo and 2D Object Tracking

Kushal Vyas
D J Sanghvi College of Engineering, University of Mumbai
Mumbai, India
Email Id: kushal.vyas@djsce.edu.in

Jonathan Joshi
Eduvance
Mumbai, India
Email Id:jon@eduvance.in

***Abstract* – Tracking of objects from a video stream is a current research challenge with widespread applications. This paper explores an approach that deals with efficient object tracking in both two dimension as well as stereo setups. It deals with real time situation involving environmental conditions, and also handling occlusion in both 2D and 3D environments. This below provides a novel way to aid in object tracking with a stereo camera pair as well as a single camera.**

Keywords – Background Subtraction, CAMshift, Depth, HSV, Kalman Filter, Motion, Occlusion, Stereo vision, Tracking.

I. INTRODUCTION

Object tracking is one of the prime application of Computer Vision. Its uses being in tracking motion of objects, surveillance, and by extension, estimating position of object as well not only in two dimensions but also in the three dimensional system.

Object tracking is a real time unsolved problem in the field of Computer Vision.Apart from environment lighting conditions and the object itself, parameter of occlusion plays a major role in the procedure of tracking. Tracking an object through all these parameters is what accounts for an efficient tracking algorithm. Various methods have been developed based on the research conducted through the years, however, obtaining a state of unambiguous tracking, as replicated by the human eyes has not yet been accomplished. We therefore rely on complex mathematical models to interpret the motion of the object in consideration. The proposed method is robust and as well as handles occlusion that arises while performing visual object tracking.

II.LITERATURE SURVEY

Over the years many methods have been developed to solve the problem- the simplest being Binary Tracking. Although it is the simplest, this method is quite unreliable in real time conditions. The first and only step is to apply a threshold, wherein we apply a pixel mask against each pixel and check whether it lies in the value set for tracking. If so, that pixel has to be tracked. In this method, it is difficult to differentiate between the actual object and another object in the background having the similar color scheme. Not to mention that the computation time for going through an image of NxN pixels will be O(n^2). Since this approach heavily relies on the lighting conditions, it provides very dubious results.

The same RGBBinary tracking can be modified by simply changing the colorspace from RGB to HSV. It's more beneficial as HSV separates the intensity and color values. So in order to track, we can use methods like histogram comparison or even by applying a mask. Since intensity depends upon lighting conditions, using HSV makes it less sensitive if not invariant to the surrounding lighting conditions [1].

Following is a motion flow approach developed by Lucas-Kanade, which worked under the assumption that the displacement of the neighboring pixels do rarely change and hence the motion vectors can be easily calculated by finding the net flow of pixels through the frames. Apart from these, advanced tracking methods have been developed,like Meanshift and on extending it, CAMshift. Instead of manually finding thresholding values, we use histogram matching techniques, so as to easily segment out the object from the scene.The approach behind the Meanshift algorithm is to find the region having the maximum pixel density based on the object histogramback projection. The disadvantage being is that it doesn't account for the direction of the object in consideration [2]. Hence this problem was solved using motion vectors while tracking. For the fact that Meanshift gave proper tracking with the simple drawback that object displacement, scaling and rotation was not considered into play, an improvement over Meanshift called ,CAMshift was used. CAMshift is a continuously adapting Meanshift algorithm wherein the search window dynamically updates itself based on the previous frame and the direction of the motion and also updates with orientation [3]. It applies Meanshift first, and once it converges, it updates the search window size as [3],

$$S = 2 \times \sqrt{\frac{M00}{256}}$$

It also calculates the orientation of the best fitting polygon into it. Furthermore, this continuously happens till the required accuracy is met.Using the concept of dynamic search windows, it saves quite some time instead of looking for the object that other tracking algorithms usually do.

 ISSN 0973-7529; ISBN 978-93-80544-20-5

Going through all the possible existing methods, the CAMshift algorithm is being used as the base algorithm with which the system is being implemented.

III. PROPOSED METHODOLOGY

The proposed method is one that is capable of tracking the object in the given system, that is a controlled environment, and also yielding the distance and depth measurements while tracking the object in real time. The approach developed is quite robust in the sense that it can easily avoid the background noises and activity and also handles occlusion of object in both 2D and stereo setups. The base algorithm being used is the CAMshift algorithm, for the reason that for an object having uniform color distribution it gives precise and robust tracking [4]. However, for multiple similar objects, an arbitrary number of histograms can be used to differentiate amongst them [4]. Thus CAMshift can be modified based on the surroundings of the object.

The setup of the system, along with camera specification can be found in the Appendix.The experiment proceeds in three broad stages, as shown below.

Fig 1. States undergone in proposed methodology

The pre-tracking and the initialization state deals with all the necessary computations that need to be carried out in order to attain a constant tracking state. It encompasses the environmental conditions, hardware calibration and tuning. The tracking state is where the actual tracking algorithm is implemented. It is here that all the possible disturbances and phenomenon such as occlusion are taken care of, before moving to the measurement state where the real time displacements are being measured.

A.Pre-Tracking and Initializations

The pre-tracking state can be further viewed in Fig 2.The camera needs to be calibrated before using it to make any real time measurements.The calibration is performed based on [5] and [6] along with a chessboard.It estimates the intrinsic and extrinsic parameters of each camerawhich can be used to undistort the image. Moving from the 2D image system to the stereo setup, the relative pose between the cameras must be known.

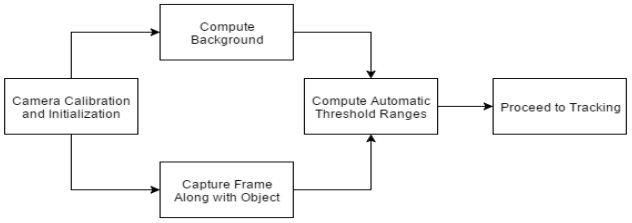

Fig 2. States undergone before Tracking

Fig 3 . Calibration, Undistortion and Rectified (left and right) image respectively

Hence stereo calibration needs to be performed, thereby rectifying the images, which facilitates in triangulation.
Background acquiring is the next major step to be done. A frame of the controlled background can be used as a reference frame for any of the difference imaging that is to be used. For such a controlled environment, background subtraction can be one of the most powerful methods, in order to get a quick estimate of where the object actually is. The outlook being that since difference imaging is used, any motion change in the image will lead to some disturbance whereas the still features will get subtracted. Various methods like frame differencing, mean filter, Gaussian average, and background mixture models [7] have been developed which aid the given problem.For this application, this was the most suitable method due to its easeof implementation as well as low computational complexity [8].Frame differencing on intersection with thresholding shows the object and for cases of multiple disturbances in the surrounding, the object can easily be narrowed down upon as per the threshold. Using that data, object position can be calculated.

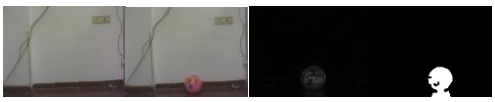

Fig 4. Background Image, Along with Object, Difference Image and Thresholded Binary Image respectively

The color extracted which is used for tracking purposes along with the initial centroid coordinates. Thisworks asan automated threshold mask.

B. Tracking State

Object can be tracked by simply tracking its centroid. A prepared HSV mask is overlapped with the newly captured frame. Such an approach is better than thresholding or binary tracking as overlapping of the mask with the frame yields the region of interest, thereby avoiding further searching inside the image. Another positive is that if the intersection of mask and frame in itself is a null image, then it can be safely concluded that the object has undergone occlusion and directly can be handled at the point of frame capture, thus not wasting more time in finding the object by expanding the search window. If the algorithm cannot predict the occluded portion, then the entire window is reset.

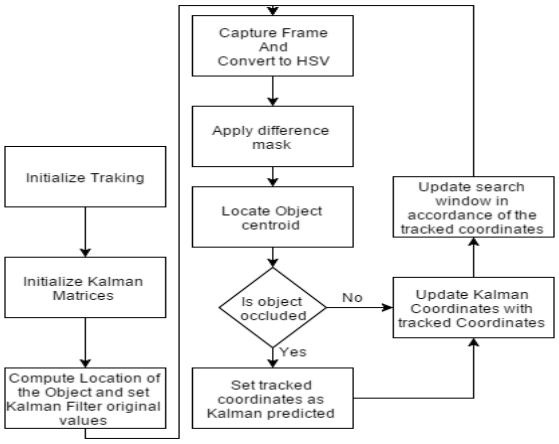

Fig 5. Process flow for Tracking

As we can see above, the entire tracking can be visualized as a process diagram.The initialization state is where the background and object detection computations have already been carried out (much in the previous parts) and the system is now ready to capture the frames on which it may apply the image transformations. Initially the search window for CAMshift is set to the full frame size and the Kalman positions are set to original centroid positions and the velocity is set to zero.The moment any frame is obtained, first it undergoes undistortion using the pre computed values of the distortion parameters of the camera. Further, an HSV mask in compliance with the difference frame is used, so as to get the object location. Thisis much similar to the pre-tracking state. Having applied the mask, it returns the probability image, which corresponds to the maximum pixel density which is the input being provided to the CAMshift Algorithm. The probability image, along with the current search window, which has been initialized to the full image resolution, return the track able centroid coordinates along with the updated search window. The search window, is where the object is to be searched in. If cases of occlusion, where the search window cannot be updated or the other way round, where the object cannot be predicted, it is re-initialized.

C. Use of Kalman Filter

The Kalman filter is an algorithm that estimatesthe state of the system based on the series of measurement observed over time [9]. It broadly has two component.

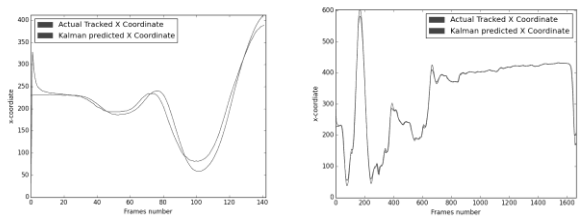

Fig 6. Tracking X Coordinate along with Kalman Filter Predictions in Two Scenarios.

One being the measurement component and the other being update. The filter designed for the setup assumes a constant-velocity model and hence considers the position and velocity in x and y axis. On applying the filer to simple 2D tracked coordinates, it gives quite an apt prediction. Kalman filter can be modified to suit accelerated motion as well.

The Kalman filter was used to continuously predict the next state of the object for every frame. Had the object got occluded fully or partially, the predicted values of Kalman filter will be used and also updated.

IV.STEREO TRACKING

The tracking method proposed above leads to a much robust image tracking, but in two dimensions. The same algorithm with some more modifications can be applied to a stereo setup as well. Tracking in 3D requires, a basic setup of stereo camera. The system was initially developed with a baseline of 20cm. The setup, mentioned in Appendix, has been calibrated and tuned to return accurate measurements, when operated within its optimum range.It is cheap alternativeover other available custom sets of stereo camera.

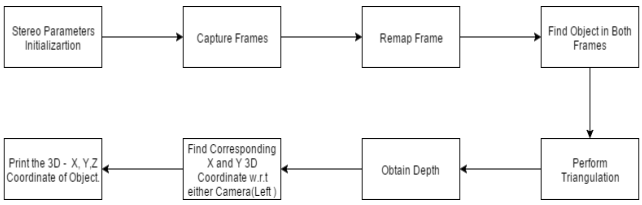

Fig 7. States for Stereo Tracking and measurement

The initialization is nothing but the loading of the precomputed stereo parameters and the matrices used for the image rectification. For every right and left camera frame that is acquired, they need to be remapped, i.e. both the frames have to be made coplanar [10]and the feature points have to lie on the same epipolar line so that the triangulation can be easily and accurately performed. We use the above mentioned method to get the tracked center positions of the object and then we apply the triangulation in order to find the depth (z coordinate) from the two images.To compute the depth, we assume the standard pinhole camera modelas shown in Fig. 8 and we obtain these equations of stereo triangulations [11].

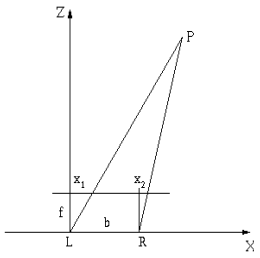

Fig 8. Computing Depth using two pinhole camera model

$$Z = \frac{BASELINE\ x\ FOCAL\ LENGTH}{DISPARITY}$$

$$X = \frac{Z\ x\ (X_l - C_l)}{FOCAL\ LENGTH}$$

$$Y = \frac{Z\ x\ (Y_l - C_l)}{FOCAL\ LENGTH}$$

$$where\ DISPARITY = X_l - X_r$$
$$C_l\ is\ Camera\ Center$$

Fig 9. Calculation of 3D World Coordinates w.r.t. Left Camera

Now as the distance of the object moves far from the camera, the readings will get more and more disruptive [10]. Because, of factors such as low resolution, and assuming one common convergence point present in both the image frames, the disparity will be zero and in that case, the depth comes out to be as infinity.

V.HANDLING OCCLUSION

Occlusion means hiding of an object in the given frame thus making it undetectable. In the system, we are using the Kalman filter to predict the values of the next state of the object when it has undergone occlusion. In two dimensional tracking occlusion is much of pixel area oriented phenomena, but when it comes to stereo tracking, handling occlusion in both ways, i.e. the pixel area paradigm as well as the depth paradigm is a real challenge. There is a very high chance that the algorithm can get confused between the two options of occlusion that are occurring.

A. Ambiguity in Occlusion in Stereo Tracking

There may be a point where in the object might be getting covered partially by some neighboring object. In such case we must get the actual position of the centroid of the object and not the half cut one. There are quite a few methods with which we can avoid such a case. One of them being, checking the area change in each frame. This will give us an idea whether the object is being cut or no. But the problem in this case is that if the object is simply moving in the Z-plane, then it'll mistake the object for the same condition.Another method is that we monitor the depth along with checking the area. This leads to a more profound decision making for the object being partially occluded or it being traversing in the Z plane.

B.Differential Based Approach for Occlusion Resolution

To solve the above problem, a differential based approach can be used. Let's say that for the past *n* frames, the object follows a certain distribution for depth and a certain one for area changes.So we keep finding the rate of change of area and the rate of change of depth at all frames. The one being higher will tell us whether the object is moving in the Z plane or being partially occluded or both.

C. Predicting Displaced Centroid for Partial Occlusion

If the object is moving not in the Z plane, then along with the Kalman filter, another custom prediction needs to be added in order to account for the apparent displacement of the centroid due to visual barriers. The object's actual centroid may be somewhere behind the obstruction whereas since a part of the object is visible, the apparent centroid is located at the center of the visible portion. Using simple Newton's laws of motions, the values of velocities and pixel area ratios can be used to reconstruct the predicted and displaced position of the actual centroid.

VI.RESULTS AND ANALYSIS

To validate the proposed methodology, few sets of readings have been taken covering all the aspects such as difference frames, depth measurement, and occlusion and so on. The system specification used for testing and runtime analysis of the algorithm, has been put down in the Appendix.

Proceeding towards the results, during the 2D tracking of the objects, the following cases were considered.

A. Raw Tracking

The modified CAMshift algorithm, on its own, show efficient object tracking. The upside being, that if the object goes out of screen, the custom algorithm will pause as the object is undetectable in the scene, whereas with the traditional implementation of the CAMshift algorithm, it would pick up the object having next nearest histogram that was of the object, thus giving erroneous results.

For instance, if the object was just occluded in a frame, the entire tracing, will be lost to the surrounding objects having a much similar histogram to that of the object. It'll be quite arduous to track the same object again. Hence the modified approach, at least eliminates such background dependencies, thereby avoiding such conditions.

B. Tracking and Occlusion Handling

The given video sequence causes the object to be occluded as well as it has been resolved using the rate of change of area method mentioned above.

Fig 10. Video Sequence for testing Occlusion

As seen, the object has been occluded in some frames as it traverses from one end to another, This has been handled using rate of change of pixel area, wherein a tolerance of 70 % was set for the object to be considered as occluded. We'll be analyzing the x coordinate of the object.

As seen in Fig. 11, the object, when starts to undergo partial occlusion, the slope of the lines tend to decrease. This suggests the apparent shift in the center of the object. The center is still predicted to be in the direction of the motion, but because the visible area has decreased, it tends to predict an inaccurate displacement. On being completely hidden, the object center cannot be found and hence it is reset to zero. The Kalman filter, being a state dependent, updates itself one frame after. When

the object partially reappears, the tracked coordinates caters to noise as well, and updates the motion.

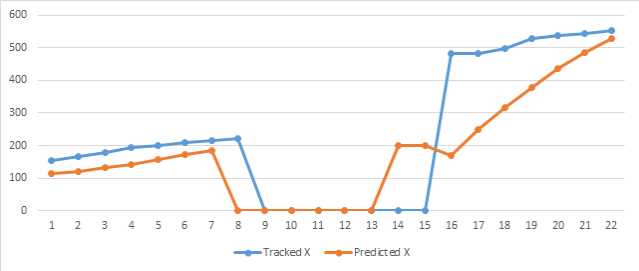

Fig 11.Handling Occlusion (Without considering shape derivative)

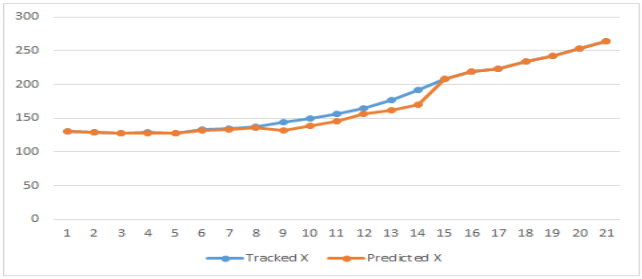

Fig 12. Handling Occlusion(With shape derivative)

When the area of object is taken under consideration as well as the velocity with which the object is traversing, the center can be predicted with much greater accuracy, even when it is completely invisible. The center can be predicted using an average velocity when it is visible, and then applying Newton's motion laws, where the time will be calculated with respect to the frame rate. As seen statistically, there is a significant amount of accuracy that is being obtained under this method. The problem of un-traceableness being observed before appears to be mitigated.

C. Distance Measurement using Stereo Setup

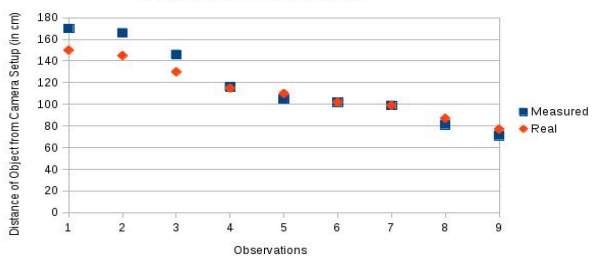

Fig 13. Depth Plot I (Real and Measured usingsetup)

The distance of a point is measured using Triangulation. Applying stereo triangulation on rectified images, the depth can be calculated and with extension its' world 3D coordinates. Few parameters such as stereo camera's with identical focal length and lenses, good resolution of the camera, the baseline distance and establishing proper stereo correspondence between the images is a must to get accurate results. The system comprising of webcams lacked the resolution criteria. [10]There were slight variations in the focal length, which may account for the slightly varied results. Apart from this, since the tracking involves tracking of the object center only, it is very important to establish correspondence amongst the images. For the given system, following was the depth accuracy obtained.

From the above figure, it is seen that the system function at an accurate and optimum level in the ranges 60cm – 120cm.A minimum of 50 cm is need to mark common convergence between the frames.Beyond 140 cm, the system cannot perform accurate triangulation. This is because, the lack of good resolution and as the object keep going far, the disparity reduces due to which even a small change in the disparity can lead to major measurement changes. This can be improved by using a camera with a much higher resolution.
Few more sets of reading were taken to verify the measurements. Consistency in output patterns support the above proposed statement.

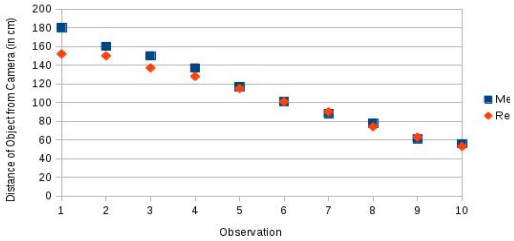

Fig 14. Depth Plot II (Real and Measured using setup)

Further analysis of the measurements, state that computing a much dense correspondence between the two frames, may increase the accuracy. Since there is a high probability that the object center in both the frames might not be the actual geometric center, due to which there may be discrepancy in the result. Apart from these, there is data loss in the image itself, since it is being returned in a JPEG format. Upon error calculation of the first and second depth plot, 3.6 % error is obtained during distance measurement for the camera in its optimum range.

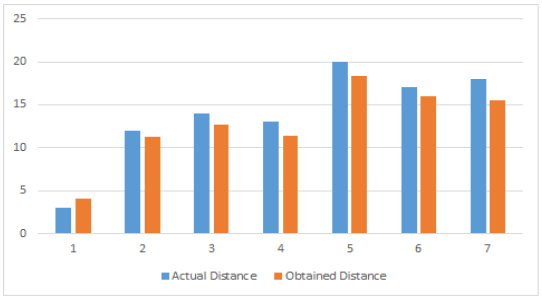

Fig 15. Actual and Obtained Real World X Coordinate Values

D. Error reduction and Improvements

To reduce error in measurements, the system may be modified with higher resolution as well as establishing stereo correspondence plays a very important role during triangulation. Apart from this, the rays generating from the object to the cameras may be skew in nature. To enhance triangulation, the Optimal Triangulation Technique [12] can be applied.

VII. CONCLUSION AND FUTURE SCOPE

This paper described a stereo as well as a 2D system and an algorithm that can lead to much efficient tracking and is suitable for obtaining robust tracking in a fixed environmental setups. It also handles occlusion quite well and is not easily susceptible to disturbances. The measurements obtained are, to quite an extent, accurate.There is much scope on pursuing this research further, with the use of high standard industry camera's which can offer greater resolution and much accurate measurements, up to a far range. Advanced computing concepts can be used for object recognition, thus eliminating the need for thresholding. Training the same algorithm to understand the disturbances due to lighting conditions and that also deal with occlusion at a much higher level is another aspect.Such algorithms can be used for surveillance and defense purposes. The algorithm can be extended to tracking multiple objects simultaneously.With usage of multiple cameras not only can the depth be measured, but also it can be an attempt to create precise structure of objects.

VIII. APPENDIX

A.System Setup Specifications

The given setup included the Logitech C - 170H computer USB Webcam offering VGA quality video capturing. The baseline distance was about 20 cm. On calibration, the focal length was around 700px. Algorithm implemented on Linux, Ubuntu 14.04 using C++ and Python programming language.

IX.ACKNOWLEDGEMENT

I would like to give a formal acknowledgement to Mr. Karan M Nair, who has been essential in teaching me research methodology and guiding me throughout the process of documenting and analyzing experiment results.

REFERENCES

[1]. Patrick Sebastian, Yap Vooi Voon, and Richard Comley, "Color Space Effect on Tracking in Video Surveillance", *International Journal on Electrical Engineering and Informatics*, Vol. 2, pp. 298 – 312, November 10,2010.

[2]. TIAN Gang, HU Rui-Min, and WANG Zhong-Yuan, ZHU Li, "Object Tracking Algorithm Based on Meanshift Algorithm Combining with Motion Vector Analysis", *First International Workshop on Education Technology and Computer Science,*Vol. 1, pp. 987 – 990, March 2009.

[3]. Bradski, G. R., "Real Time face and object tracking as a component of a perceptual user interface", Fourth IEEE Workshop on Applications of Computer Vision, WACV'98, 1998.

[4]. D. Exner, E. Bruns, D. Kurz, A. Grundhofer and O. Bimber, 'Fast and robust CAMShift tracking', *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010.

[5]. Birchfield, S. and Tomasi, C., "*A pixel dissimilarity measure that is insensitive to image sampling*", IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998.

[6]. Z. Zhang, "A flexible new technique for camera calibration", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330-1334, 2000.

[7]. M. Piccardi, "Background subtraction techniques: a review", *2004 IEEE International Conference on Systems, Man and Cybernetics*, 2004.

[8]. E. Komagal, A. Vinodhini, Archana and Bricilla, "Real time Background Subtraction techniques for detection of moving objects in video surveillance system", *2012 International Conference on Computing, Communication and Applications*, 2012.

[9]. R. Faragher, "Understanding the Basis of the Kalman Filter Via a Simple and Intuitive Derivation [Lecture Notes]", *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 128-132, 2012.

[10]. P. Bagga, 'Real Time Depth Computation Using Stereo Imaging', *JEEE*, vol. 1, no. 2, p. 51, 2013.

[11]. Linda Shapiro and George Stockman – *Computer Vision*; 2003, p. 431.

[12]. R. Hartley and A. Zisserman– *Multiple view geometry in computer vision*; Cambridge, UK: Cambridge University Press, 2003, p. 318.