This full text paper was peer-reviewed at the direction of IEEE Instrumentation and Measurement Society prior to the acceptance and publication.

Depth Estimation in Still Images and Videos Using a Motionless Monocular Camera

Sotirios Diamantas, Stefanos Astaras, Aristodemos Pnevmatikakis
Multimodal Signal Analytics
Athens Information Technology
44 Kifisias Avenue, 15125 Marousi
Athens, Greece
Email: {sodi, sast, apne}ait.gr

***Abstract*—In this research we address the problem of depth estimation using a *single* motionless monocular camera. In our method we make no use of reference objects or marks in the image plane or on the ground apart from a one-off object used for horizon line detection; even this, however, is not necessary if a vanishing point detection algorithm is employed. Camera height is the only known parameter that is projected onto the image plane. Our algorithm has been tested using both a light calibrated and a non-calibrated camera and the results presented demonstrate that it works exceptionally well with both options. Our method promises to relax several assumptions and restrictions followed by state-of-the-art methods such as the height or width of the object of interest. Furthermore, our algorithm has been tested on still images as well as on videos using a background subtraction algorithm for automatic segmentation of foreground moving objects. The results obtained demonstrate our method is accurate and useful to a variety of applications from robot navigation to target tracking.**

I. Introduction

Depth estimation has long been one of the most important fields of research in the domains of robotics and computer vision. For decades researchers have been approaching the problem of depth estimation using a variety of sensors and/or methodologies. Depth estimation can become a less complex problem should it be approached through Time-of-Flight (TOF) sensors such as laser range scanners. However, TOF sensors alone cannot provide enough information about the environment, the objects, and the surroundings. Therefore, a combination of TOF sensors, mainly, laser range scanners, and cameras or stereo camera systems have so far been used for tackling the problem of depth estimation. Stereo vision on the other hand, has been the focus of a large number of researchers. While it has been a heavily studied problem, stereo vision requires corresponding two or more frames which is a non-trivial task. The occlusion of objects within different views, the change in illuminance and decalibration in a stereo vision system make correspondence a challenging and a far from solved problem.

In this research the problem of depth estimation is addressed using a single still camera. Our method aims at relaxing assumptions made by state-of-the-art methods and promises to provide a basis for tackling a number of complex problems such as object and scene recognition, 3D reconstruction, localization and mapping, and robot navigation. Furthermore, our research on monocular depth estimation surpasses any problems and complexity emanating from corresponding different views while at the same time makes no assumptions about the environment, the objects it consists of, or any known landmarks or marks on the image plane apart from an initial one-off estimation of horizon line using a known object that is removed in all subsequent perceived images even when the environment changes from outdoor to indoor. Our algorithm has been tested on still images manually segmented of the object of interest as well as on videos using a foreground/background algorithm for automatic segmentation of the object of interest. In addition, our algorithm has been evaluated on a light calibrated camera as well as on a non-calibrated camera.

This paper consists of five sections. Section II is devoted to background and related work on depth estimation. In Section III we present the methodology followed for tackling the problem of depth estimation using a monocular camera. Section IV presents the results from our method and, finally Section V provides a discussion on the conclusions drawn from this research as well as a discussion on future research.

II. Background Work

The majority of works that tackle the problem of depth estimation deal with stereo (binocular) vision. A stereo vision algorithm with applications in obstacle avoidance is presented in [1]. In [2] an overview is given of the taxonomy of dense two-frame correspondence algorithms based on their performance. They are categorized onto local or global methods with the former referring to window-based methods favoring performance over accuracy, while the latter are favoring accuracy over performance. 3D reconstruction has been of particular interest the past few years, especially with the immense popularity of digital cameras and smartphones. In [3], [4] the authors exploit the motion of cameras to densely reconstruct objects and environments using multi-view geometry principles. In [5] a depth estimation method is presented using a monocular camera along with a supervised learning method by taking into account the global structure of the scene. In this paper the authors have collected a set of outdoor images with ground-truth depth maps used to train their model and then apply supervised learning to predict the depth map as a function of the image. Learning depth from monocular

978-1-5090-1817-8/16/$31.00 ©2016 IEEE

images using deep convolutional neural fields appears in [6]. In that research the authors propose a deep structured learning scheme which learns the unary and pairwise potentials of the continuous conditional random field in a unified deep neural network framework. In [7] the authors present various methods for estimating heights of objects using vanishing lines from the ground as well as vertical vanishing point using an uncalibrated camera. In their method a reference object with known height is used in the image plane. In [8] a method for measuring the heights of any feature based on an uncalibrated camera is presented. Their research is based on the Focus of Expansion (FOE) under pure translation.

Depth estimation with a monocular camera and by means of optical flow and least squares is presented in [9]–[11]. In this work the authors have trained the system with a large number of varying optical flow vector magnitudes with respect to a range of speeds. A regression analysis has been used for the purpose of estimating depth based on a formula derived from regression analysis. A least squares strategy has been employed by taking snapshots of a landmark at various positions and is compared against optical flow [9]. A method that performs fusion of the two strategies, namely optical flow and least squares appears in [10]. A depth estimation method by means of externally calibrating a single camera appears in [12].

The MOG algorithm [13] is a well-known background subtraction method that utilizes mixtures of Gaussians. The concept of using adaptive mixture models continues to provide solutions for various applications. [14] and [15] are popular implementations of the mixture-of-Gaussians algorithm, tuned to real-world applications. [16] uses nonparametric kernels instead of Gaussian functions, and [17] uses the local binary pattern operator, to create texture models instead of color models. [18] and [19] use pixel color sample models, instead of estimating a probability density function. [20] and [21] fuse the color model with the texture model and the contour model (shape), respectively. There are also solutions outside the concept of adaptive pixel models; [22] uses neural networks, while [23] uses a fuzzy system.

III. Methodology

We propose a new method to achieve depth estimation from monocular images or video. In our method, the camera remains fixed on a tripod and it captures both still images as well as video sequences of moving objects. For our depth estimation algorithm to work there are two light assumptions that need to be considered:

- The input to the algorithm is the height of the camera from the ground. This is the case for cameras that remain fixed either on a mobile unit such as a mobile robot or are attached onto a pole such as surveillance cameras.
- The object whose depth is to be computed has to be either fully visible in the camera -in case horizon line lies above the contour of the object- or be visible from the horizon line to the point the object touches the ground -in case the horizon line crosses the object. If the object is on top of another object then this object has to be visible at the point it touches the ground.

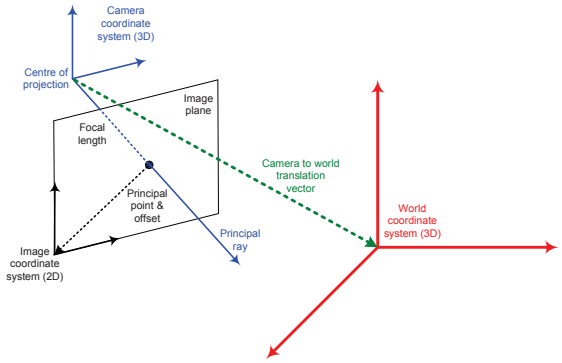


Fig. 1. Pictorial representation of 3D world onto a 2D image plane and their respective coordinate systems.

The orientation of the camera relative to the world can be described using two 3D Cartesian coordinate systems: that of the world and that of the camera. The camera coordinate system has its origin at the center of projection of the camera, and its z_c axis along the principal ray of the camera. Figure 1 shows the projection of 3D world into 2D image plane.

As it was mentioned earlier, the only input to our algorithm is the known camera height from the ground, that is in m, H_{C_m}. The camera height is then projected onto the image plane in order to derive the camera height in pixels H_{C_p}. The projected camera height on the image plane is the vector between the horizon line and the lower part of the object that touches the ground. H_{O_m} and H_{O_p} are the dimensions of the known object in meters and in pixels, correspondingly. The following eqn. (1) expresses the relationship between camera and object height from which we derive the horizon line and the projected camera height. The first step required for depth estimation is to compute the horizon line. This can be done in two ways. The first is to utilize a vanishing point algorithm. This approach is particularly useful when dealing with indoor environments where geometric objects appear and thus extracting lines and hence vanishing lines is a rather easy task. For the horizon line we are only interested in the $y-axis$ component of the vanishing point. The second approach which by the way has been followed in this research is to make use of an object whose height is known. This is a one-off case and the object need not be present at all instances or along with the object whose depth is to be computed.

$$\frac{H_{C_m}}{H_{C_p}} = \frac{H_{O_m}}{H_{O_p}} \tag{1}$$

Upon obtaining H_{C_p}, the horizon line can be determined in the image plane. For estimating the depth to an object, a known dimension of the object, e.g., height, is normally required. However, in our algorithm this is not necessary because the projected camera height onto the image plane can act as an imaginary reference object without having to have a visible

reference object in the image plane thus adding to the accuracy of the algorithm that entails the circumvention of segmenting an entire object. In other words, the bounding box enclosing the object of interest need not faithfully represent the entire segmented object so long as the lower part of the object is accurately enclosed by the bounding box.

The next step is to estimate depth. Depth, d, is computed using the following eqn. 2

$$d = \frac{f_{mm} H_{O_m} H_{Img_p}}{H_{O_p} H_{S_{mm}}} \tag{2}$$

where f_{mm} is the focal length in camera in mm, H_{O_m} is the camera height in meters, that is the only known parameter, H_{Img_p} is the height of the image in pixels, H_{O_p} the height of the object in pixels, and $H_{S_{mm}}$ the sensor size in mm.

For obtaining the focal length of the camera, f_{mm}, this can either be carried out through intrinsic camera calibration or be obtained from the lens of the camera. In our method, we have obtained results using both a calibrated and an uncalibrated camera. For our experiments we used a SONY Cybershot TX9 camera with 12.2 MP image resolution. Object height in pixels can easily be obtained from the image either by manually segmenting the object of interest or by automatically segmenting the object in motion as is described in the forthcoming subsection.

A. Background Subtraction

In order to separate the important objects in each frame from their surroundings, we employ a background subtraction algorithm. In the simplest case, the background is constant, so it is memorized, and pixels that are different must belong to the foreground objects. In a real-world application, the background pixels sustain slow or repetitive variations, due to environmental factors (sunlight variations, wind, shadows). A good background subtraction algorithm should be able to discriminate between these expected variations and abrupt changes that would denote a foreground object.

The background subtraction algorithm that we have used is a nonparametric kernel method (KDE) based on [16], as implemented in the algorithmic library *bgslibrary* [24]. The key concept is that the probability density function of the background pixel intensities is estimated, as more samples are collected, by a weighted sum of kernel functions, each centered on a sample. This method adapts to sunlight changes, as new samples update the model so that it follows the actual density function. Small displacements of the camera or other background objects due to the wind are accounted for, both by the background model (if the displacements are periodical) and by observing the neighborhoods of the pixels; a shadow detection scheme has also been employed.

To improve the result, we apply morphological filters to the resulting foreground mask (opening and closing), which reduce noisy false positives while trying to preserve the original shape of the objects. Foreground pixels are then organized into connected components, in order to localize the separate foreground objects, and calculate their bounding boxes. For human detection, we use an opening matrix that is taller than it is wide, so that the effect of false negatives separating a silhouette to more than one connected component (e.g. at the neck or the legs) is prevented; this leaves possible noisy flickering unaffected, due to its random nature. In the end, by observing the calculated depth and height of each bounding box, we can dismiss those that appear to be noisy instead of actual foreground objects (e.g. too far and tall, or too short and close).

IV. Results

In this section the results from the depth estimation method are presented. Experiments have been carried out both in indoor and outdoor environments, using still images as well as videos, and making use of a light calibrated and a non-calibrated camera. Figure 2 depicts a human subject at varying depths in both indoor and outdoor environments. In this figure, the depth estimation algorithm has been tested on still images. In the first figure, Fig. 2 (R), an object whose height is known is depicted; this is used only once and the object is not needed in the image plane along with the object whose distance from camera is to be estimated as is the case in [7]. In Figure 2 (a) the human object is at a distance of 7m from the camera and the estimated distance is 6.99m and 7.34m, for the calibrated and uncalibrated method, respectively (a relative error of 0.14% and 4.86%, respectively). Similarly, in Fig. 2 (b) the subject is at a distance of 14m from the camera and the estimated distance is 14.08m and 14.78m, for the calibrated and uncalibrated method (a relative error of 0.57% and 5.57%, respectively). In the subsequent figure, i.e., Fig. 2 (c) the human subject is 7m further from the previous figure, i.e., at a distance of 21m. Our algorithm has computed the depth being equal to 20.75m and 21.77m, that is 1.19% and 3.67% for the calibrated and uncalibrated scenarios, respectively. In the last outdoor figure, i.e., Fig. 2 (d) the human object is at a ground-truth distance of 28m. The estimated depth is 27.16m (3%) and 28.5m (1.79%) for the calibrated and uncalibrated method, respectively. In Figures 2 (e)-(f) the environment has switched to indoor and the horizon line need not to be recalculated since camera height from the ground is fixed. In Figure 2 (e) the ground-truth distance of the human object is at 5m from the camera and the estimated distance is 5.12m (2.4%) and 5.38m (7.6%) for the calibrated and uncalibrated method, respectively. Similarly, in Fig. 2 (f) the distance of the human subject from the camera is at 10m while the estimated is 10.39m (3.89%) and 10.91m (9.06%) for the calibrated and uncalibrated method, respectively. Finally, Fig. 2 (g) the actual distance between the subject and the camera is 13m with the estimated being at 13.77m (5.92%) and 14.46m (11.23%) for the calibrated and uncalibrated methods, respectively. Table I summarizes the results from the depth estimation algorithm. Columns 3 - 5 demonstrate the results derived from a calibrated and an uncalibrated camera. The mean error of the calibrated camera is 2.44% whereas the mean error of the uncalibrated camera

is 6.25%, that is a 3.8% difference between the calibrated and uncalibrated methods.

TABLE I
DEPTH ESTIMATION RESULTS WITH STILL IMAGES - CALIBRATED (C) / UNCALIBRATED (U) METHOD

Fig.	Act. Depth	Est. Depth (C/U)	Abs. Error	Rel. Error (%)
2(a)	7	6.99 / 7.34	0.01 / 0.34	0.14 / 4.86
2(b)	14	14.08 / 14.78	0.08 / 0.78	0.57 / 5.57
2(c)	21	20.75 / 21.77	0.25 / 0.77	1.19 / 3.67
2(d)	28	27.16 / 28.5	0.84 / 0.50	3.00 / 1.79
2(e)	5	5.12 / 5.38	0.12 / 0.38	2.40 / 7.60
2(f)	10	10.39 / 10.91	0.39 / 0.91	3.89 / 9.06
2(g)	13	13.77 / 14.46	0.77 / 1.46	5.92 / 11.23

Figure 3 depicts a human subject at varying distances within an indoor as well as an outdoor environment. In this figure, a method for automatic background subtraction in videos has been employed [16]. The camera's focal length has been computed from a known object's dimension. The first four figures depict a human subject in an outdoor environment at distances which are multiples of 7 (meters). The last three figures depict a human subject within an indoor environment at distances of 5, 10m, and 13m. The relative error obtained in both environments vary between 0.2% and 1.68% with the exception of the first figure where the error is 10.43%; this can be considered an outlier case. The mean relative error is as low as 2.26%; it is 0.9% for the inlier case. Table II summarizes the results obtained using a foreground/background subtraction algorithm for segmenting a human subject [1].

TABLE II
DEPTH ESTIMATION RESULTS WITH VIDEO - AUTOMATIC BACKGROUND SUBTRACTION

Fig.	Act. Depth	Est. Depth	Abs. Error	Rel. Error (%)
3(a)	7	6.27	0.73	10.43
3(b)	14	13.85	0.15	1.07
3(c)	21	21.11	0.11	0.52
3(d)	28	28.47	0.47	1.68
3(e)	5	4.99	0.01	0.2
3(f)	10	9.93	0.07	0.7
3(g)	13	12.84	0.16	1.23

V. CONCLUSIONS AND FUTURE WORK

In this research we have addressed the problem of depth estimation using minimal information about the camera and the environment. In particular, we have managed to estimate depth between a camera and an object at varying distances and environments using a *single* non-moving monocular camera with camera height from the ground being the only known parameter. The focal length of the camera can be obtained from the lens specifications of the camera or the metadata of the image (uncalibrated method). A number of experiments have been carried out in indoor as well as in outdoor environments using still images and videos with a human subject in motion. The results obtained from our method show the robustness of our approach both in the manual object selection and the automatic object selection methods. In particular, the mean relative error is almost identical in both scenarios (2.4% and 2.3% for still images and automatic video background subtraction methods, respectively).

This research comes at a point where visual measurements, such as Google's Project Tango [25], are essential in smartphones and other handheld devices for providing precise measurements, area learning, and augmented reality among others. Moreover, our method is essential for mobile robot navigation in unknown and unstructured environments. As a future work we will be extending our method to objects that are partly visible or partly occluded. Furthermore, we will be implementing shape from shadow techniques for estimating an object's height and depth.

[1] A depth estimation video appears in our lab's video repository: https://www.youtube.com/user/AITSmartLab

REFERENCES

[1] A. G. Lazaros Nalpantidis, Ioannis Kostavelis, "Stereovision-based algorithm for obstacle avoidance," in *Intelligent Robotics and Applications*, vol. 5928, 2009, pp. 195–204.

[2] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[3] M. Pizzoli, C. Forster, and D. Scaramuzza, "REMODE: Probabilistic, monocular dense reconstruction in real time," in *IEEE International Conference on Robotics and Automation*, 2014, pp. 2609–2616.

[4] G. Vogiatzis and C. Hernandez, "Video-based, real-time multi-view stereo," *Image and Vision Computing*, vol. 29, no. 7, pp. 434–441, 2011.

[5] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. C. Platt, Eds. MIT Press, 2006, pp. 1161–1168. [Online]. Available: http://papers.nips.cc/paper/2921-learning-depth-from-single-monocular-images.pdf

[6] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, 2015.

[7] A. Criminisi, I. Reid, and A. Zisserman, "Single view metrology," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 123–148, 2000.

[8] Z. Chen, N. Pears, and B. Liang, "A method of visual metrology from uncalibrated images," *Pattern Recognition Letters*, vol. 27, no. 13, pp. 1447–1456, 2006.

[9] S. C. Diamantas, A. Oikonomidis, and R. M. Crowder, "Depth estimation for autonomous robot navigation: A comparative approach," in *International Conference on Imaging Systems and Techniques*, Thessaloniki, Greece, 2010, pp. 426–430.

[10] ——, "Depth computation using optical flow and least squares," in *IEEE/SICE International Symposium on System Integration*, Sendai, Japan, 2010, pp. 7–12.

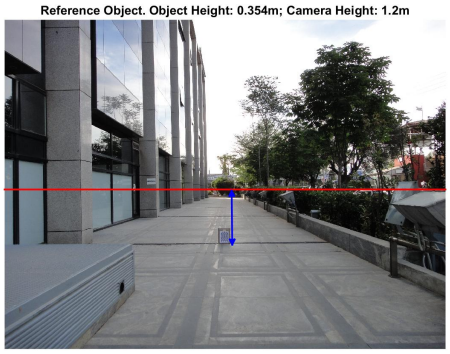
[11] S. C. Diamantas, "Biological and metric maps applied to robot homing," Ph.D. dissertation, School of Electronics and Computer Science, University of Southampton, 2010.

[12] G. Bardas, S. Astaras, S. Diamantas, and A. Pnevmatikakis, "3d tracking and classification system using a monocular camera," *Wireless Personal Communications, Accepted*, pp. 1–23.

[13] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *1999 Conference on Computer Vision and Pattern Recognition (CVPR '99), 23-25 June 1999, Ft. Collins, CO, USA*, 1999, pp. 2246–2252. [Online]. Available: http://dx.doi.org/10.1109/CVPR.1999.784637

[14] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Video-Based Surveillance Systems*. Springer US, 2002, ch. 11, pp. pp 135–144.

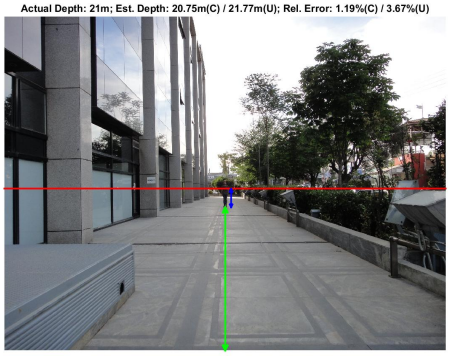
[15] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, August 23-26, 2004.*, 2004, pp. 28–31. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2004.1333992



(R)

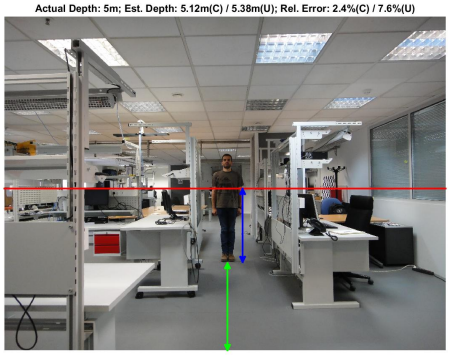
(a)

(b)

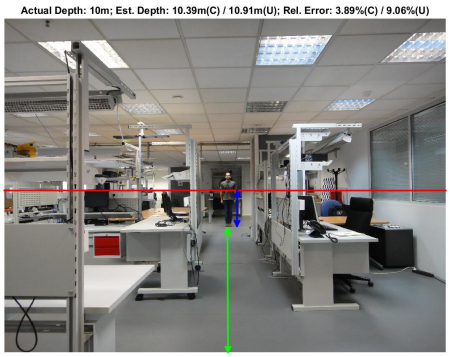


(c)

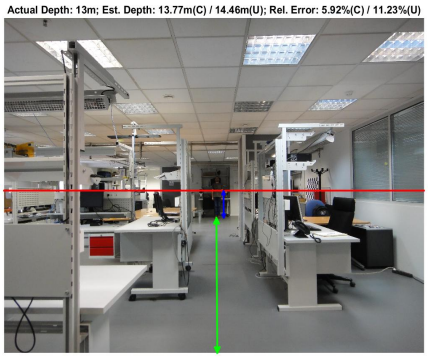
(d)



(e)



(f)



(g)

Fig. 2. (R) A one-off reference object used for horizon line detection. (a) - (g) Figures depict a human subject in still images at varying distances and environments (indoor-outdoor). Red line denotes the horizon line whereas blue and green arrows denote projection of the camera height onto the image plane and the depth to be estimated between the camera and the human subject, respectively. In figure captions (C) and (U) denote results obtained from a calibrated and an uncalibrated method, respectively.

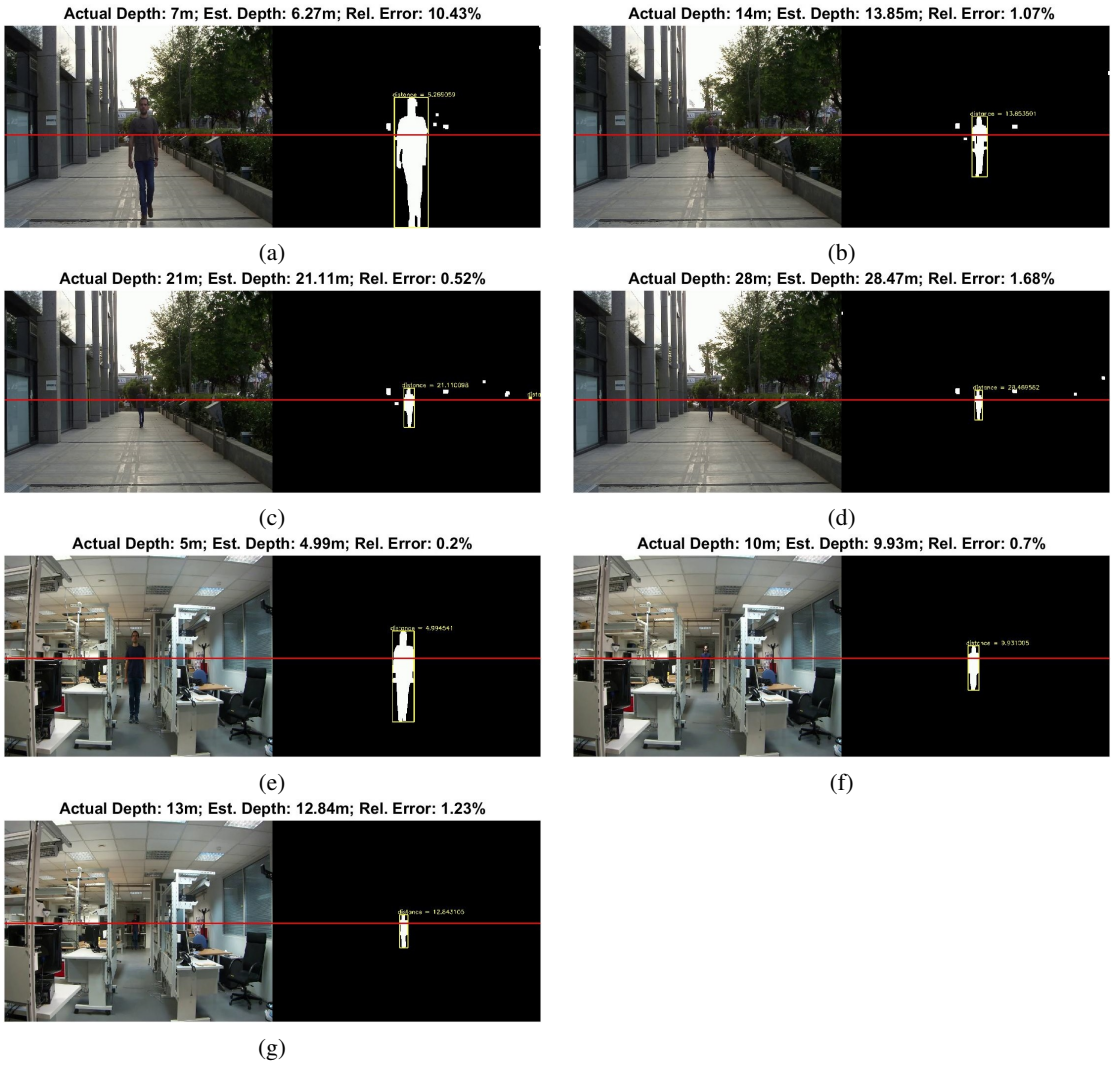


Fig. 3. (a) - (g) Figures depict a human subject at varying distances both in an indoor and an outdoor environment. In this set of figures, a background subtraction algorithm has been employed for automatic moving object subtraction. The red line denotes the horizon line.

[16] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," in *Proceeding of the IEEE*, vol. 90, no. 7, November 2002, pp. 1151–1163.

[17] M. Heikkilä and M. Pietikäinen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 657–662, 2006. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TPAMI.2006.68

[18] O. Barnich and M. V. Droogenbroeck, "Vibe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011. [Online]. Available: http://dx.doi.org/10.1109/TIP.2010.2101613

[19] A. B. Godbehere, A. Matsukawa, and K. Y. Goldberg, "Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation," in *American Control Conference, ACC 2012, Montreal, QC, Canada, June 27-29, 2012*, 2012, pp. 4305–4312.

[20] J. Yao and J.-M. Odobez, "Multi-layer background subtraction based on color and texture," in *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*, 2007. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2007.383497

[21] S. Noh and M. Jeon, "A new framework for background subtraction using multiple cues," in *Computer Vision - ACCV 2012 - 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part III*, 2012, pp. 493–506.

[22] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1168–1177, 2008. [Online]. Available: http://dx.doi.org/10.1109/TIP.2008.924285

[23] ——, "A fuzzy spatial coherence-based approach to background/foreground separation for moving object detection," *Neural Computing and Applications*, vol. 19, no. 2, pp. 179–186, 2010. [Online]. Available: http://dx.doi.org/10.1007/s00521-009-0285-8

[24] A. Sobral, "BGSLibrary: An OpenCV C++ Background Subtraction Library," in *IX Workshop de Viso Computacional (WVC'2013)*, Rio de Janeiro, Brazil, Jun 2013.

[25] Project Tango, "https://www.google.com/atap/project-tango/," 2016.