Efficient Vehicle Counting and Classification using Robust Multi-Cue Consecutive Frame Subtraction

By CH. Manaswini

JNTUK, India

Abstract- The ability to count and classify vehicles provides valuable information to road network managers, highways agencies and traffic operators alike, enabling them to manage traffic and to plan future development of the network. Increased computational speed of processors has enabled application of vision technology in several fields such as: Industrial automation, Video security, transportation and automotive. The proposed method in this paper is a robust adaptive multi-cue frame subtraction method that detects foreground pixels corresponding to moving and stopped vehicles, even with noisy images due to compression. First the approach adaptively thresholds a combination of luminance and chromaticity disparity maps between the learned background and the current frame. The segmentation is further used by a two-step tracking approach, which combines the simplicity of a linear 2-D Kalman filter and the complexity of 3-D volume estimation using Markov chain Monte Carlo (MCMC) methods. The experimental results shows that the proposed method can count and classify vehicles in real time with a high level of performance under challenging situations, such as with moving casted shadows on sunny days, headlight reflections on the road using only a single standard camera.

Keywords: *3D volume estimation, tracking, vision technologies.*

GJCST-F Classification: *I.4.0*

Efficient Vehicle Counting and Classification using Robust Multi-Cue Consecutive Frame Subtraction

CH. Manaswini

Abstract- The ability to count and classify vehicles provides valuable information to road network managers, highways agencies and traffic operators alike, enabling them to manage traffic and to plan future development of the network. Increased computational speed of processors has enabled application of vision technology in several fields such as: Industrial automation, Video security, transportation and automotive. The proposed method in this paper is a robust adaptive multi-cue frame subtraction method that detects foreground pixels corresponding to moving and stopped vehicles, even with noisy images due to compression. First the approach adaptively thresholds a combination of luminance and chromaticity disparity maps between the learned background and the current frame. The segmentation is further used by a two-step tracking approach, which combines the simplicity of a linear 2-D Kalman filter and the complexity of 3-D volume estimation using Markov chain Monte Carlo (MCMC) methods. The experimental results shows that the proposed method can count and classify vehicles in real time with a high level of performance under challenging situations, such as with moving casted shadows on sunny days, headlight reflections on the road using only a single standard camera.

Keywords: 3D volume estimation, tracking, vision technologies.

I. Introduction

The continuous increase in the congestion level on public roads, especially at rush hours, is a critical problem and a challenging situation for vehicle tracking used to count and classify vehicles. The existing methods for traffic management, surveillance and control are not adequately efficient in terms of the performance, cost and the effort needed for maintenance and support. Vision-based video monitoring systems offer a number of advantages for traffic flow surveillance combine information from cameras with other Intrusive technologies [1]. Though it has some complexities, computer vision can be used to obtain richer information, such as analyzing the visual features apart from geometry of the vehicle. Background subtraction forms an important component in many of these applications. The central idea behind this representation is that can then be utilized for the classification of a new observation.

The information provided by such a module can then be considered as a valuable low-level visual cue to perform high-level object analysis tasks such as object detection, tracking, classification and event analysis.

II. Related Work

The main contributions of the proposed approach are background subtraction and the tracking of vehicle.

a) Background Subtraction

A procedure used in video image processing to subtract away persistent noise patterns generated in the video camera or the optical system. Also used to display motion as changes between the current and old "background" video images Background subtraction is a computational vision process of extracting foreground objects in a particular scene. A foreground object can be described as an object of attention which helps in reducing the amount of data to be processed as well as provide important information to the task under consideration. Often, the foreground object can be thought of as a coherently moving object in a scene. We must emphasize the word coherent here because if a person is walking in front of moving leaves, the person forms the foreground object while leaves though having motion associated with them are considered background due to its repetitive behavior. In some cases, distance of the moving object also forms a basis for it to be considered a background, e.g. if in a scene one person is close to the camera while there is a person far away in background, in this case the nearby person is considered as foreground while the person far away is ignored due to its small size and the lack of information that it provides. Identifying moving objects from a video sequence is a fundamental and critical task in many computer-vision applications. A common approach is to perform background subtraction, which identifies moving objects from the portion of video frame that differs from the background model. Methods that model the variation of the intensity values of

Author: Department of ECE, V R Siddhartha Engineering College.
e-mail: manaswini1323@gmail.com

Figure 1 : Sample of Vehicles with Shadows in Sunny Daylight

background pixel. **[2]Unimodal Distribution:** Unimodel distributions play an important role in background subtraction modeling scheme and gives satisfactory classification rate. They are fast and simple but are not able to adapt to multiple backgrounds, e.g., when there are trees moving in the wind. *Drawbacks:* This approach works only under static background but not in dynamic background scenario. **[3]The mixture of Gaussian:** In compensation to the unimodel distributions to handle dynamic background most researchers proposed several works on adaptive background modeling approach. Sophisticated adaptation methods are required to solve major two problems in dynamic scene: changes of illumination and changes of background content. Among several approaches proposed Mixture of Gaussian (MOG) model can solve several problems occurred in dynamic background, especially in case of repetitive background motion such as waving tree. MOG and other methods proposed update background models using linear model which is not adequately adapted according to the changes in the background scene. *Drawbacks:* They cannot handle fast variations with accuracy using a few Gaussians, and therefore, this method has problems for the sensitive detection of foreground regions. **[4]The nonparametric kernel density estimation:** This model uses pixel intensity as the basic feature for modeling the background. This model keeps a sample of intensity values for each pixel in the image and uses this sample to estimate the density function of the pixel intensity distribution, therefore able to estimate the probability of any newly observed intensity value. This model can handle the situations where the background of the scene is cluttered and not completely static but contain small motions. The model is updated continuously and adapt to the changes in scene background. *Drawbacks:* This method is time consuming because pixel based technique assume that the time series of observation is independent of each pixel. **[5]Codebook based methods:** The most sensitive color-based background subtraction methods, applied both indoors and outdoors scenarios, even with some motion in the background. Codebook (CB) background subtraction algorithm was intended to sample values

over long times, without making parametric assumptions. The key features of our algorithm are in the followings: (1) Resistance to artifacts of acquisition, digitization and compression,(2) Capability of coping with illumination changes,(3) Adaptive and compressed background models that can capture structural background motion over a long period of time under limited memory, (4) Unconstrained training that allows moving foreground objects in the scene during the initial training period. The CB algorithm adopts a quantization/clustering technique, to construct a background model. Samples at each pixel are clustered into the set of code words. The background is encoded on a pixel by pixel basis. *Drawback:* Codebook method does not evaluate probabilities, which is computationally very expensive. Approaches using **Gradient cues:[6]**Use of gradient cues instead of intensity values improves the robustness against illumination changes. However, plain regions that may be present in some vehicles are not extracted and still need further processing for discriminating shadows. Shadows in the real world belong to so-called global illumination effects, because the light ray on its way from the light source to the camera is affected by more than only one refection on an object surface. In Fig. 1, the formation of a cast shadow is shown. The light coming from a single light source reaches the background only partially due to a moving object. The darkened region on the background is called cast shadow. **Umbra** is the shadow that receives illumination coming only from diffuse ambient light, **Penumbra** receives illumination from both the ambient light and a portion of direct light. Penumbra has more chromatic similarity with respect to its original color than in the case of umbra.

According to the taxonomy proposed in **[7]**, shadow suppression methods can be classified as **deterministic** and **statistical**. **Deterministic methods** uses on/off decision processes It can be subdivided into **Model** and **Non model** based. **Model-based methods:** Model-based method requires more acknowledge about environment and are computationally more demanding. They use explicit models of the vehicles to be tracked and also of light sources. *Drawbacks* in model-based deterministic techniques can obtain better results for shadow suppression, but it should be remarked that these are excessively cumbersome for a practical implementation in outdoor traffic surveillance. **Nonmodel-based methods** do not use explicit models of the vehicles to be tracked and also of light sources. Advantages of Nonmodel based deterministic approaches are most suitable for outdoors applications **Statistical models:** A probability function is used when decision making for different membership groups (is shadow or not). This method can be subdivided into **parametric** and **nonparametric**. **Parametric** approaches use a series of parameters that determine the characteristics of the statistical functions of the model

Nonparametric approaches automate the selection of the model parameters as a function of the observed data during training. Advantage nonparametric statistical methods are best for indoors, since the scene is more constant, and thus their statistical description is more effective. More recently in[8], it is shown that the improved hue, luminance, and saturation (IHLS) color space is better suited for change detection and shadow suppression than HSV and normalized red, green, and blue (RGB).
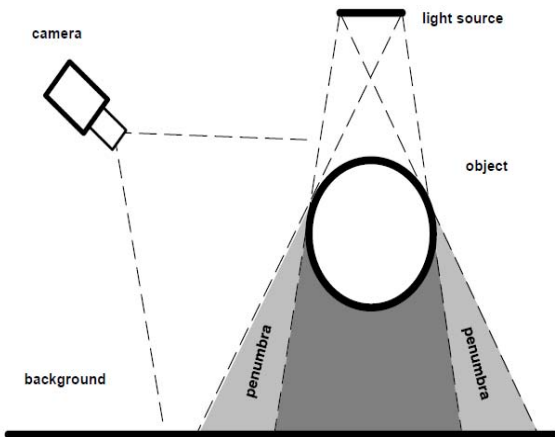

Figure 2 : Cast shadow generation: The scene grabbed by a camera consists of a moving object and a moving cast shadow on the background. The shadow is caused by a light source of certain extent and exhibits a penumbra

III. Multicue Frame Subtraction

A simple method of subtracting one movie frame from another will provide information about which parts of the scene have changed (generally due to motion). This method was performed on each frame of the movie, with consecutive frames being subtracted from each other. First, the scene is converted to an array of pixel values. These pixel values are the averaged **Red, Green, and Blue (RGB)** values for each pixel. The pixel values of the previous frame are then subtracted from the current frame's pixel values, and the absolute value of the values is taken. The result is an array of values that represent how much each pixel has changed between the two frames, with higher values representing more change. The amount of change in a region of pixels can be interpreted as the amount of motion that is taking place in that region. These data can then be used to determine where in the scene the most motion is taking place. Multiple cues play a crucial role in image interpretation. A vision system that combines shape, colour, motion, prior scene knowledge and object motion behavior is described. Grimson et al [6] expound on the benefits of using multiple cues to disambiguate complex scenes. They use depth from stereo and colour as attentional mechanisms to identify candidate regions in the image which are then subjected to detailed shape analysis for object recognition purposes. In our recent work in computer vision we have also embarked on an approach which exploits multiple cues to reduce the computational complexity of image interpretation to a manageable level[9] [10] .

The term "cue" is understood in a very general sense and can mean also prior knowledge or contextual information whether spatial or temporal. Our approach differs from other attempts to exploit multiple cues not only in the extent but also in the scope of cues used and the flexibility of their combination. In other words, cues are not combined in a preprogrammed (fixed) way but rather their combination depends on the goal of interpretation, on the image content and on the stage of interpretation. All three are dynamically changing. In addition to the usual image properties such as shape, motion and colour, our interpretation scheme is unique. Fusing different cues has been proven to be the current best approach to obtain accurate background subtr-action results. Two main issues arise at this point: 1) which cues are to be used and 2) in which way they are fused. Typical useful cues are pixel color, pixel intensity (gray level), and edges (obtained from image gradients). The sequential steps to be given for the proposed Multicue Frame Subtraction. In this algorithm, I is the current image ($Iy\ xy$, $Icx\ xy$, and $Icy\ xy$ are the luminance and chromaticity coordinates for each xy pixel in the IHLS color space); B is the mean background ($By\ xy$, $Bcx\ xy$, and $Bcy\ xy$ are the luminance and chromaticity coordinates for each xy pixel in IHLS color space); ΣI is the background luminance variance; Σc is the background chromaticity variance; kl and kc are the proportionality constants for determining the background/ foreground disparity maps (Dl and Dc); ol, oc, os, and ow are the offsets applied for pixel classification; and finally, tg and tf are the thresholds.
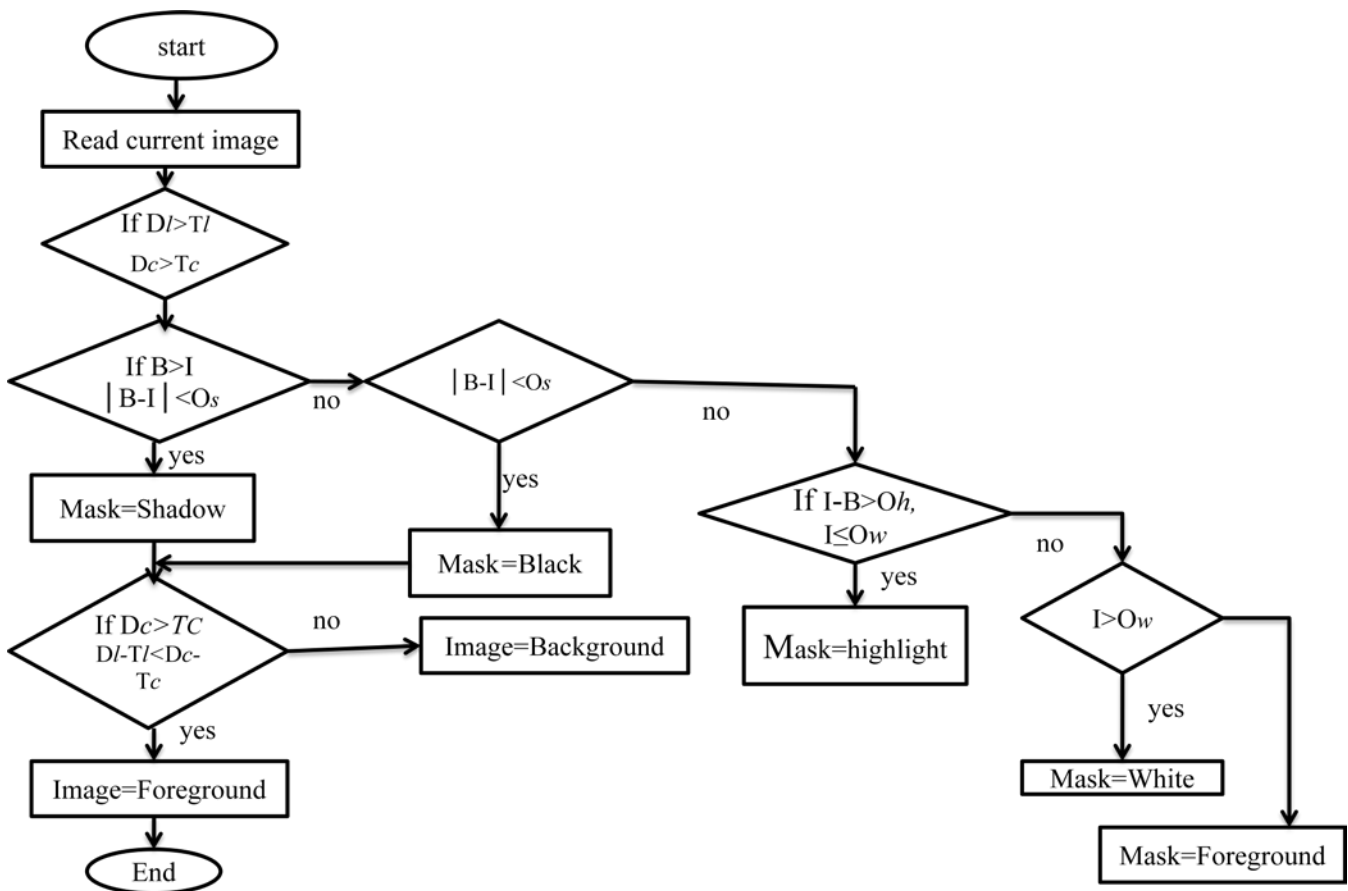
Figure 3 : Multi-cue Frame Subtraction Flowchart

The novelty of our approach relies on a Multi-cue Frame Subtraction procedure in which the segmentation thresholds can adapt robustly to illumination changes, maintaining a high sensitivity level to new incoming foreground objects and effectively removing moving casted shadows and headlight reflections on the road. A tracking module provides the required spatial and temporal coherence for the classification of vehicles, which first generates 2-D estimations of the silhouette of the vehicles and then augments the observations to 3-D vehicle volumes by means of a Markov chain Monte Carlo (MCMC) method.

IV. HIGHLIGHT CROPPING

The mask will be processed to remove highlighted regions corresponding to sudden illumination changes due to weather variability or vehicle headlights. Additionally, the blob mask in between are filled by applying the watershed procedure [11] to *foreground*, *highlight* and *white* in between, filling them with *foreground*. Ignoring the darker regions to build vehicle blob candidates for the tracking stage, but highlighted regions require further processing to remove those generated by sudden illumination changes coming from weather variations or headlights. Taking into account lane geometry, where x is the direction along the lane, and y is the transversal in the images these directions match with x and y of the image. Inside a lane, the pixels corresponding to a line perpendicular to the lane are summed, and if all non background pixels correspond to the *highlight* category, then those pixels are set to *background*. Vehicle projections usually have more than one pixel category in lines perpendicular to lanes, and therefore, using this approach, blobs can be cropped per lane, removing fully highlighted areas but not the vehicles. Lane geometry is considered in this process as there can be vehicles in different lanes, parallel to highlighted regions, which could interfere in the cropping.

V. VEHICLE TRACKING

The Multicue mask is used as the observation of the tracking process that estimates the position and volume of vehicles. Tracking is carried out in a two-step process that first obtains 2-D bounding boxes of the projection of the vehicles in the road plane and then estimates their 3-D volume according to the calibration of the camera[12].

a) Camera Calibration

From a mathematical point of view, an image is a projection of a three dimensional space onto a two dimensional space. Geometric camera calibration is the

process of determining the 2D-3D mapping between the camera and the world coordinate system [13]. Therefore, obtaining the taken with fixed internal parameters; the rigidity of the scene provides sufficient information to recover calibration parameters.

Regarding 2-D tracking schemes, the Kalman filter has been shown to offer great estimation results in rectified images, where the dynamics of the vehicles are undistorted and thus can be assumed to be linear 2-D estimations lack the required accuracy in classification strategies: The viewpoint of the camera is a critical aspect for these strategies. The perspective effect is reduced, and the length and width of vehicles can be measured with lower error. However, more flexible approaches should consider several potential road viewing angles. This issue directly

Figure 4 : Highlight cropping algorithm

affects the maximum accuracy that a 2-D approach can provide, and only 3-D methods can reliably determine vehicle measurements in such situations.

b) 3D Volume Estimation

[13]The proposed solution is based on a Markov Chain Monte Carlo (MCMC) method, which models the problem as a dynamic system and naturally integrates the different types of information into a common mathematical framework. This method requires the definition of a sampling strategy, and the involved density functions (namely, the likelihood function and the prior models). Typically, the complexities of these kind of sampling strategies are too high to run in real time MCMC methods have been successfully applied to different nature tracking problems They can be used as a tool to obtain maximum a posteriori (MAP) estimates provided likelihood and prior models. Basically, MCMC methods define a Markov chain, $\{\mathbf{x}_i\ t\}\ Ns\ i=1$, over the space of states, $\mathbf{x}$, such that the stationary distribution of the chain is equal to the target posterior distribution $p(\mathbf{x}_t\ |Z_t\ )$. A MAP, or point estimate, of the posterior distribution can be then selected as any statistic of the sample set (e.g. sample mean or robust mean), or as the sample, $\mathbf{x}_i\ t$, with highest $p(\mathbf{x}_{it}\ |Z_t\ )$, which will provide the MAP solution to the estimation problem. The analytical expression of the posterior density can be decomposed using the Bayes' rule as:

$$p(\mathbf{x}_t\ |Z_t\ ) = kp(\mathbf{z}_t\ |\mathbf{x}_t)p(\mathbf{x}_t\ |Z_t-1) \qquad (1)$$

Where $p(\mathbf{z}_t\ |\mathbf{x}_t)$ is the likelihood function that models how likely the measurement $\mathbf{z}_t$ would be observed given the system state vector $\mathbf{x}_t$, and $p(\mathbf{x}_t\ |Z_t-1)$ is the prediction information, since it provides all the information we know about the current state before the new observation is available. The constant k is a scale factor that ensures that the density integrates to one. For each image, the observation is the current 2D silhouette of the vehicle projected into the rectified image [13].

Considering the cuboid-model of the vehicle, and that the yaw angle is approximately zero we can reproject a 3D ray from the far-most corner of the projected cuboid and the optical center. the likelihood function must be any function that fosters volume hypotheses near the reprojection ray. For the sake of simplicity, we choose a normal distribution on the point-line distance. The covariance of the distribution expresses our confidence about the measurement of the 2D silhouette and the calibration information. The likelihood function can be written as

$$p(z_t/\mathbf{x}_t) \propto \exp(\ (yt-xt)^{\mathrm{T}}s\text{-}1(yt\text{-}xt))$$

Where $\mathbf{x}_t$ is a volume hypothesis, and $\mathbf{y}_t$ is its projection onto the reprojection ray. The position of $\mathbf{y}_t$ can be computed from $\mathbf{x}_t$ as the intersection of the ray
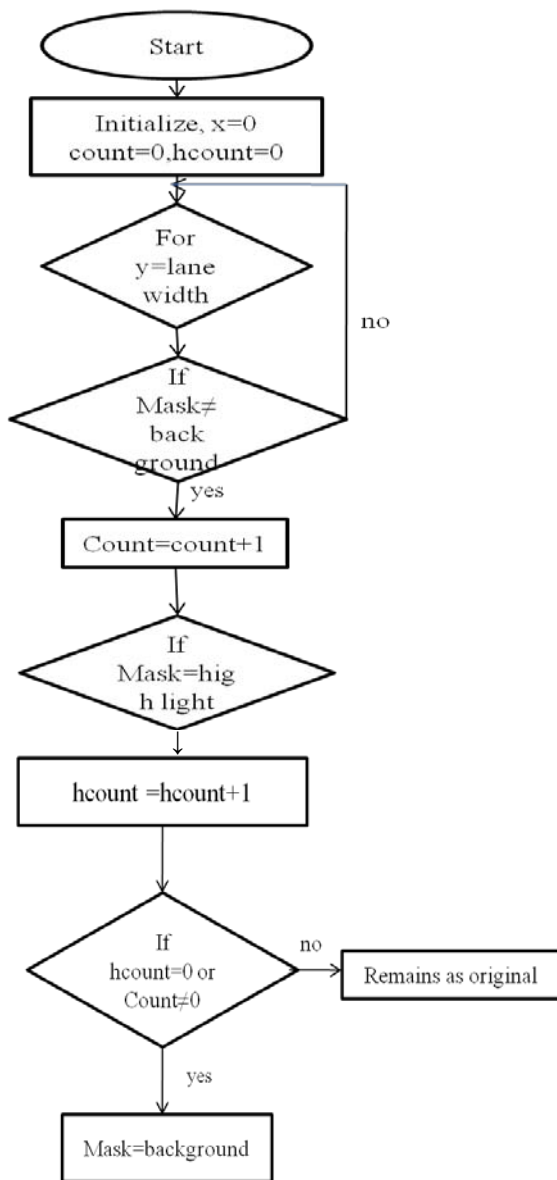
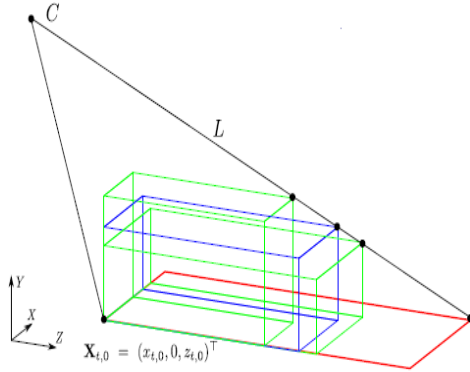and a plane passing through $\mathbf{x}t$ and whose normal vector is parallel to the ray. For this

Figure 5 : Projective ambiguity: a given 2D observation in the OXZ plane (in red) of a true 3D cuboid (blue)may also be the result of the projection of a family of cuboids (in green) with respect to camera C

purpose we can represent the ray as a Plücker matrix Lt = $\mathbf{ab}^{\mathrm{T}} - \mathbf{ba}^{\mathrm{T}}$, where $\mathbf{a}$ and $\mathbf{b}$ are two points of the line, e.g. the far-most point of the 2D silhouette, and the optical center, respectively. These two points are expressed in the *WHL* coordinate system (i.e.,width, length, height). Therefore, provided that we have the calibration of the camera, we need a reference point in the 2D silhouette. We have observed that the point with less distortion is typically the closest point of the quadrilateral to the optical center, whose coordinates are $\mathbf{X}t,0 = (xt,\mathbf{o},0, zt,0)^{\mathrm{T}}$ in the *XYZ* world coordinate system. This way, any *XYZ* point can be transformed into a *WHL* point as $\mathbf{x}t = R_o\mathbf{X}t - \mathbf{X}t,_o$. Nevertheless, the relative rotation between these systems can be approximated to the identity, since the vehicles typically drive parallel to the *OZ* axis. The plane is defined as Πt = $(\mathrm{n}t^T, Dt)^{\mathrm{T}}$, where $\mathrm{n}t = (n_x, ny, nz)^{\mathrm{T}}$ is the normal to the ray Lt, and $Dt = -\mathrm{n}t^T\mathbf{x}t$. Therefore, the projection of the point on the ray can be computed as $\mathbf{y}t = Lt\Pi t$.

VI. VEHICLE COUNTING AND CLASSIFICATION

An images of the video used for the test is a sunny day where vehicles project shadows are taken. The duration of the video is 3min, and the traffic flow is dense with several vehicles passing in parallel. The vehicle types include cars, motorbikes, heavy trucks, articulated trucks, vans, and buses, but for this test, we consider three classes, depending on tracked region geometric characteristics: *Two Wheels*, *Light Vehicle*, and *Heavy Vehicle*. In this test, we have compared our segmentation approach with other vision-based recent alternative, maintaining the same tracking procedure i.e., Modified Codebook (MCB), This method have been adapted to our context to update the background and use adaptive luminance and chromaticity thresholds in the same way as ours, which is referred in the test as Adaptive Multicue Frame Subtraction (AMC).

The main differences between MCB with respect to the color-only part of the AMC come from the use of cylindrical RGB color space instead of IHLS and the segmented categories. In the *Sunny* sequence, the main difficulty comes from the detection of dark vehicles that project shadows. In some cases, there may be some dark vehicles that do not have sufficient gradient features to identify that they are vehicles. MCB has particular problems in distinguishing these cases as it does not take into account cues other than color. In this case, it might have seemed desirable to apply extra morphologic operations, such as erosion, but it was not appropriate because other good regions would have been lost for the test. Thus, the proposed approach for improving the segmentation by estimating the shadow direction shows its relevance in this scenario [15].

Figure 6 : Sample vehicles of the test video for classification

VII. CONCLUSION AND FUTURE SCOPE

This paper introduces a real-time method to augment 2D vehicle detections into 3D volume estimations by using prior vehicle models for efficient counting and classification of vehicles in sunny day(including passing vehicles cast shadows). This system distinguishes itself from other computer-vision-based approaches in the way in which it can handle casted shadows without the need for any hardware other than cameras, such as GPS to estimate the direction of the shadows. Hence, we believe that this is a viable alternative to replace other vision based approaches and intrusive technologies whose installation and maintenance are more cumbersome than using cameras only.

The extension to the approach can be done for vehicles in different weather conditions and tracking under severe occlusion can be achieved by adding Markov Random Field factors to the posterior distribution expression the 3-D model can understand the image projections of two or more vehicles if intersected.
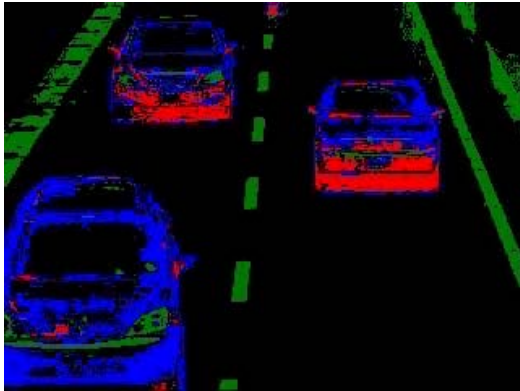
VIII. Results for the Proposed Method

Figure 7 : Resutls for Muticue Frame Subtraction
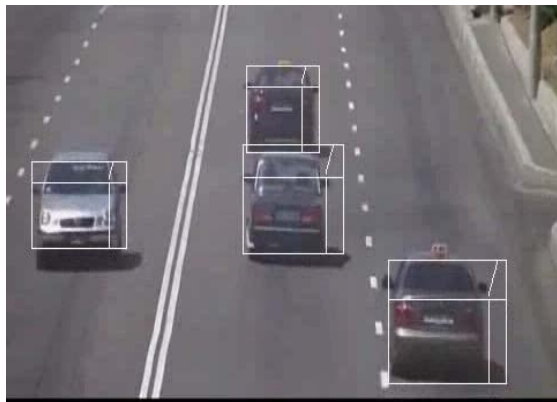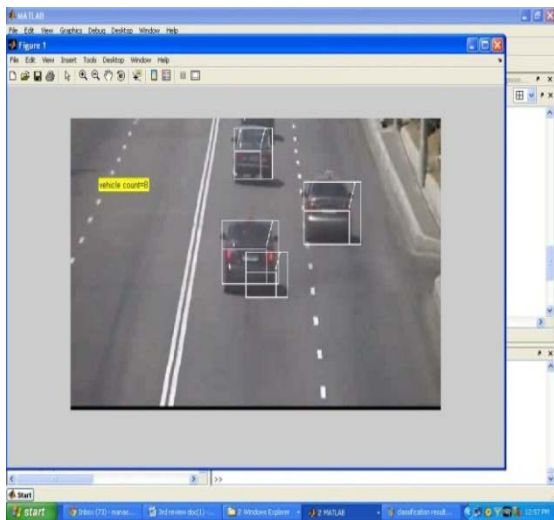

Figure 8 : Results for 3D tracking

Figure 9 : The yellow bar shows the counting results passing in the video

Classification Results for the Proposed Method.

Vehicle	Width	Length	Height
TW	20	10	17.35
LV	39	17	23.2941
HV	66	32	35

The overall performance of the proposed method is 92% when compared to the Modified codebook method (MCB) which is 84%. These results show that our proposed method yields good results compared to other vision-based alternative approaches.

References Références Referencias

1. Luis Unzueta, *Member, IEEE*, Marcos Nieto, Andoni Cortés, Javier Barandiaran, Oihana Otaegui, and Pedro Sánchez "Adaptive Multicue Background Subtraction for Robust Vehicle Counting and Classification" IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 13, NO. 2, JUNE 2012.
2. C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," IEEE Trans. Pattern Anal. Mach.Intell., vol. 19, no. 7, pp. 780–785, Jul. 1997.
3. C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, 1999, pp. 246-252.
4. AHMED ELGAMMAL, RAMANI DURAISWAMI, MEMBER, IEEE, DAVID HARWOOD, AND LARRY S. DAVIS, "Background and Foreground Modelling Using Nonparametric Kernel Density Estimation for Visual Surveillance", FELLOW PROCEEDINGS OF THE IEEE, VOL. 90, NO. 7, JULY 2002.
5. Kyungnain Kim', Thanarat H. Chalidabhongse', David Hanuood', Larry Davis', "Background Modelling and Subtraction by Codebook Construction" International Conference on Image Processing (ICIP) 0-7803-8SS4-3/04/$20.00 2004 IEEE.
6. Jurgen Stauder, Roland Mech, Jorn Ostermann. "Detection of Moving Cast Shadows for Object Segmentation" IEEE TRANSACTIONS ON MULTI MEDIA, VOL. 1, NO. 1, MARCH 1999.
7. Andrea Prati, Member, IEEE, Ivana Mikic, Member, IEEE, Mohan M. Trivedi, Member, IEEE,and Rita Cucchiara, Member, IEEE, "Detecting Moving Shadows: Algorithms and Evaluation" IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 25, NO. 7, JULY 2003.
8. J Kittler, J Matas, M Bober and L Nguyen "IMAGE INTERPRETATION: EXPLOITING MULTIPLE CUES" INTERNATIONAL CONFERENCE ON IMAGE PROCESSING AND APPLICATIONS, EDINBURGH, JUNE 3-6, 1995.
9. J Crowley and H I Christensen. *"Vision as Process".* Springer-Verlag, Berlin, 1995.
10. J. Matas, P Remagnino, J Kittler, and J Illingworth. Control of scene interpretation. In *Vision as Process*, pages 347 373.Springer-Verlag, 1994.

11. E. Preteux, "Watershed and skeleton by influence zones: A distance-based approach," *J. Math. Imag. Vis.*, vol. 1, no. 3, pp. 239–255, Sep. 1992.

12. X. Zou, D. Li, and J. Liu, "Real-time vehicles tracking based on Kalman filter in an ITS," in *Proc. Int. Symp. Photoelectron. Detection Imag.*, 2007, vol. SPIE 6623, p. 662 306.

13. Z. Khan, T. Balch, and F. Dellaert, "MCMC-based particle filtering for tracking a variable number of interacting targets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1805–1819, Nov. 2005.

14. M. Nieto, L. Unzueta, A. Cortés, J. Barandiaran, O. Otaegui, and P. Sánchez, "Real-time 3D modeling of vehicles in low-cost monocamera systems," in *Proc. Int. Conf. Comput. VISAPP*, Algarve, Portugal, Mar. 2011, pp. 459–464.

15. Z. Mayo and J. R. Tapamo, "Background subtraction survey for highway surveillance," in *Proc. Annu. Symp. PRASA*, Stellenbosch, South Africa, 2009, pp. 77–82.