International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Saliency Technique for Efficient Back Ground Subtraction

Aiswarya Muralidharan [1], S. Sivakumar[2]

[1]ME Communication system, Department of Electronics and Communication, Sri Shakthi Institute of Engineering and Technology, Coimbatore

[2]Assistant Professor Department of Electronics and Communication, Sri Shakthi Institute of Engineering and Technology, Coimbatore

Abstract: ***The key technique for automatic video analysis is Background subtraction, especially in the domain of video surveillance. In this paper Saliency detection can be used in background subtraction. The proposed method is effectively used in the case where the scene is highly changing or the camera is not moving. The separation of foreground and back ground regions within and across video frames is done by visual and motion saliency information. Input video contains all this information which can be extracted. The saliency induced features is effectively combined by conditional random field (CRF) which deals with unknown pose and scale variations of the foreground object .The method is computationally efficient, reliable, and simple to implement and thus it can be easily extended to various applications. It is effective in detecting saliency compared to various State-of-the-art methods.***

Keywords: Saliency detection, change detection, Condition random field, background subtraction, SGMM, visual saliency, motion saliency

1. Introduction

A human can easily determine the subject of interest in a video, even though that subject is presented in an unknown or cluttered background or even has never been seen before. With the complex cognitive capabilities exhibited by human brains, this process can be interpreted as simultaneous extraction of both foreground and background information from a video. Many researchers have been working toward closing the gap between human and computer vision. However, without any prior knowledge on the subject of interest or training data, it is still very challenging for computer vision algorithms to automatically extract therefore ground object of interest in a video.

The demand for detecting the automated motion and tracking of object promoted research activity in the field of computer vision. In this paper proposes a method in which visual and motion saliency is taken in to action. The background subtraction process, involves the detection of foreground pixels, labelled and they are grouped into regions by a connected components algorithm. It deals with lighting changes, repetitive motions from clutter, and long-term scene changes with different weather conditions . However, problems arise when moving objects mixed with each other and one object enters the scene while another is leaving. In addition moving shadows are not removed during tracking.

Non-rigid object tracking with a moving camera based on the mean shift algorithm .One advantage is that the tracker performance is not affected by intense blurring due to camera motion which is always a problem for contour based trackers. It also works under low quality sequences. The feature selection, a tracking algorithm based on affine change models, and a technique for monitoring features during tracking. The feature selection criterion depended entirely on how well the tracker worked. The change detection is a task used as a first step in many computer vision applications such as video surveillance, medical diagnosis or human-computer interaction. In an image sequence, our aim is to identify for each frame the set of pixels that are significantly different from the previous frames. The requirements and constraints of the detection algorithm are different for different applications. In this paper change detection has been extensively used in order to segment foreground objects from the background. Foreground objects are associated between frames in order to perform a scene analysis and detect events of interest. The parts of the scene which are normally observed are considered as background. Therefore, it is assumed that the background can be well described by means of a statistical model, the background model. Background subtraction algorithms use a model of the static scene, the background model, to distinguish between background and foreground in video sequences.

The process of segmentation of foreground objects by detecting the changes with reference to a background model is called as background subtraction.This paper proposes a method to detect and measure motion based upon tracking salient features using a model of visual attention. Natural scenes are often composed of several entities, from which usually only a small portion are relevant to tasks such as, object recognition, area surveillance, event detection, or path planning. In fact the ability to separate informative regions from the background clutter is an essential requirement to perform these assignments successfully. Biological systems have developed to be remarkably effective in focusing their visual attention to relevant targets, as opposed to computer vision where background subtraction is still an unsolved problem. Commonly background subtraction has been approached by detecting moving objects against a static background .While effective in certain scenes, this approach has severe problems when the scenes are dynamic or the camera is not static. These situations have been addressed by trying to compensate for the camera movements and by continuously updating the background model. However accurate camera movement estimation is not an easy problem and rapid background model updating is often technically difficult, if not impossible. Furthermore these methods are not applicable at all if we have a single image instead of video, or if the objects of interest are not moving against the background.In this paper, we aim at automatically extracting foreground objects in videos which are captured by freely moving cameras. Instead of assuming that the background motion is dominant and different from that of the foreground as did, we relax this assumption and allow foreground objects to be presented in freely moving scenes. We advance both visual and motion saliency information across video frames is utilized for integrating the associated features for VOE (i.e., visual saliency, shape, foreground/background colour models, and spatial/temporal energy terms). From our quantitative and qualitative experiments, we verify that our VOE performance exhibits spatial consistency and temporal continuity, and our method is shown to outperform state-of the-art unsupervised VOE approaches. It is worth noting that, our proposed VOE framework is an unsupervised approach, which does not require the prior knowledge (i.e., training data) of the object of interest nor the user interaction for any annotation.

Most existing unsupervised VOE approaches assume the foreground objects as outliers in terms of the observed motion information, so that the induced appearance, color, etc. features are utilized for distinguishing between foreground and background regions. However, these methods cannot generalize well to videos captured by freely moving cameras as discussed earlier. In this work, we propose a saliency-based VOE framework which learns saliency information in both spatial (visual) and temporal (motion) domains. By advancing conditional random fields (CRF), the integration of the resulting features can automatically identify the foreground object without the need to treat either foreground or background as outliers.

In general, one can address VOE problems using supervised or unsupervised approaches. Supervised methods require prior knowledge on the subject of interest and need to collect training data beforehand for designing the associated VOE algorithms. For example, Wu and Nevatia and Lin and Davis both decomposed an object shape model in a hierarchical way to train object part detectors, and these detectors are used to describe all possible configurations of the object of interest (e.g. pedestrians). Another type of supervised methods requires user interaction for annotating candidate foreground regions. For example, image segmentation algorithms proposed in focused on an interactive scheme and required users to manually provide the ground truth label information. Although the color features can be automatically determined from the input video, these methods still need the user to train object detectors for extracting shape or motion features. Recently, researchers proposed to use some preliminary strokes to manually select the foreground and background regions, and they utilized such information to train local classifiers to detect the foreground objects. While these works produce promising results, it might not be practical for users to manually annotate a large amount of video data.

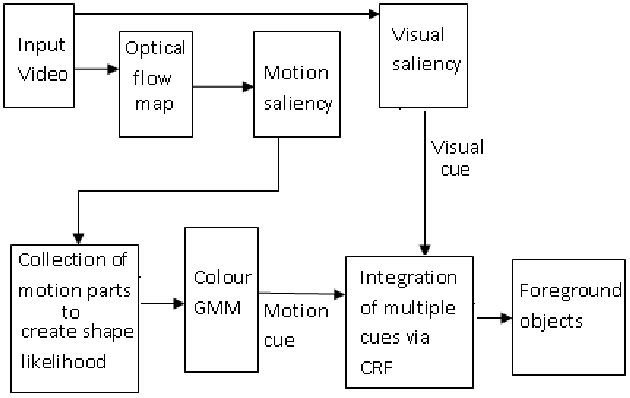
2. Existing Method

Gaussian mixture model is the probabilistic method of background subtraction. Compared to state of art method it is more adaptable and multimodal compared to state of art method and requires low memory. In statistics, a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs. Formally a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. However, while problems associated with "mixture distributions" relate to deriving the properties of the overall population from those of the sub-populations, "mixture models" are used to make statistical inferences about the properties of the sub-populations given only observations on the pooled population, without sub-population identity information. Some ways of implementing mixture models involve steps that attribute postulated sub-population-identities to individual observations (or weights towards such sub-populations), in which case these can be regarded as types of unsupervised learning or clustering procedures. However not all inference procedures involve such steps. Mixture models should not be confused with models for compositional data, i.e., data whose components are constrained to sum to a constant value (1, 100%, etc.). However, compositional models can be thought of as mixture models, where members of the population are sampled at random. Conversely, mixture models can be thought of as compositional models, where the total size of the population has been normalized to the two layered system analyses the video frame at two levels: pixel level and region level. At former level pixel classified on the result obtained by subtracting the two complementary background model and later ,new static foreground regions are classified as static or removed objects. To avoid incorporation of static foreground objects in to the background model information are fed back at the pixel level. Foreground detection is used to determine the areas of image belonging to foreground class with respect to the similarity in input frame and background model.

3. Proposed Method

The saliency of an item – be it an object, a person, a pixel, etc. – is the state or quality by which it stands out relative to its neighbours. Saliencydetection is considered to be a key attention mechanism that facilitates learning and survival by enabling organisms to focus their limited perceptual and cognitive resources on the most pertinent subset of the available sensory data.

Saliency typically arises from contrasts between items and their neighbourhood, such as a red dot surrounded by white dots, a flickering message indicator of an answering machine, or a loud noise in an otherwise quiet environment. Saliency detection is often studied in the context of the visual system, but similar mechanisms operate in other sensory systems. What is salient can be influenced by training: for example, for human subjects particular letters can become salient by training. When attention deployment is driven by salient stimuli, it is considered to be bottom-up, memory-free, and reactive. Attention can also be guided by top-down, memory-dependent, or anticipatory mechanisms, such as when looking ahead of moving objects or sideways before crossing streets. Humans and other animals have difficulty paying attention to more than one item simultaneously, so they are faced with the challenge of continuously integrating and prioritizing different bottom-up and top-down influences.



4. Video Object Extraction

A desirable video object extraction scheme for content based applications should meet the following criteria:

- Segmented object should conform to human perception i.e., semantically meaningful objects should be segmented.
- Segmentation algorithm should be efficient and achieve fast speed.
- Initialization should be simple and easy for users to operate (human intervention should be minimized). One feasible solution
- that satisfies these criteria is edge change detection.

In Video Object (VO) segmentation methods, which are using mathematical morphology and perspective motion model, objects of interest should be initially outlined by human observer. From the manually specified object boundary, the correct object boundary is calculated using a morphological segmentation tool. The obtained VOP is then automatically tracked and updated in successive frames. It has difficulty in dealing with a large non rigid object movement and in the presence of occlusion, especially in the VOP tracking schemes.

The algorithm based on edge change detection allows automatic detection of the new appearance of a VO. The edge change detection for inter-frame difference is another stream of popular schemes because it is straightforward to implement and enables automatic detection of new appearance. This ability enables to develop a fully automated object-based system, such as an object-based video surveillance system. It is found that the algorithms based on inter frame change detection render automatic detection of objects and allow larger non rigid motion compared to mathematical morphology and perspective motion model methods. The drawbacks are small false regions detected by decision error due to noise. Thus, small whole removal using morphological operation and removal of false parts like uncovered background by motion information are usually incorporated. Another drawback in edge change detection is that object boundaries are irregular in some critical image areas, which must be smoothened and adapted by spatial edge information. Since spatial edge information is useful for generating VOP with accurate boundaries, a simple binary edge difference scheme may be assumed to be a good solution. In order to overcome boundary inaccuracy multiple features, multiple frames and spatial-temporal entropy methods are used. In addition, it gives robustness to noise and occluding pixels.

The first stage is applied to the first two frames of a video shot to discover moving objects while the second stage is applied to the rest of the frames to extract the detected objects through the video shot. The algorithm is applied to first two frames of the image sequence to detect the moving objects in the video sequence. First two frames of video sequence are taken and motion vectors are computed using Adaptive Rood Pattern Search (ARPS) algorithm. Simultaneously, components of optical flow are computed for each block in the image. By using the motion vectors, motion compensated frame is generated.

Initial segmentation is performed on the first frame of traffic sequence. Applying watershed transformation directly on the gradient of image results in over segmentation. To avoid over segmentation morphological gradient is computed on the frame and then watershed transformation is performed. After watershed transformation, some regions may need to be merged because of possible over-segmentation.

Canny binary edge image is used to localize an object in subsequent frames of video sequences and to detect the true weak edges. Intensity edge pixels are used as feature points due to the key role that edges play in the human visual process and the fact that edges are little affected by variation of luminance. Object models evolve from one frame to the next, capturing the changes in the shape of objects as they move. The algorithm naturally establishes the temporal correspondence of objects throughout the video sequence, and the output of the algorithm is a sequence of binary models representing the motion and shape changes of the objects.

Object model is obtained by subtracting background edge from edge image and eliminating unlinked pixels. After a binary model for the object of interest has been derived the motion vectors generated from ARP's algorithm are used to match the subsequent frames in the sequence. Matching is performed on edge images because it is computationally efficient and fairly insensitive to changes in illumination. The degree of change in the shape of an object from one frame to the next is determined based on simplified Hausdorff distance where simplified Hausdorff distance is defined as combination of distance transformation and correlation. Distance Transform the image and then threshold it by different amounts to form different dilated image sets. To search for the object in the image, it is required to obtain the amount by which the image is dilated such that maximum points in the object model are matched to image set.

In this automatic VO segmentation algorithm edge change detection starts with edge detection which is the first and most important stage of human visual process. Edge information plays a key role in extracting the physical change of the corresponding surface in a real scene, exploiting simple difference of edges for extracting shape information of moving objects in video sequence suffers from great deal of noise even in stationary background. This is due to the fact that the random noise created in one frame is different from the one created in the successive frame, and thus results in slight changes of the edge locations in the successive frames. Thus difference edge of frames suppresses the noise in luminance difference by means of canny edge detector.

5. Visual Saliency

Visual attention may be a solution to the inability to fully process all locations in parallel. However, this solution produces a problem. If you are only going to process one region or object at a time, how do you select that target of attention? Visual salience helps your brain achieve reasonably efficient selection. Early stages of visual processing give rise to a distinct subjective perceptual quality which makes some stimuli stand out from among other items or locations. Our brain has evolved to rapidly compute salience in an automatic manner and in real-time over the entire visual field. Visual attention is then attracted towards salient visual locations.

The core of visual salience is a bottom-up, stimulus-driven signal that announces "this location is sufficiently different from its surroundings to be worthy of your attention". This *bottom-up* deployment of attention towards salient locations can be strongly modulated or even sometimes overridden by *top-down*, user-driven factors. Thus, a lone red object in a green field will be salient and will attract attention in a bottom-up manner (see illustration below). In addition, if you are looking through a child's toy bin for a red plastic dragon, amidst plastic objects of many vivid colors, no one color may be especially salient until your top-down desire to find the red object renders all red objects, whether dragons or not, more salient.

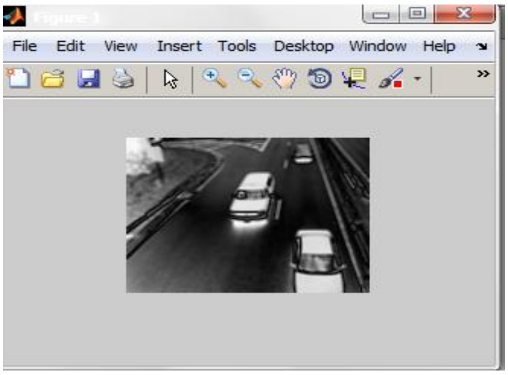


Figure 2: Visual Salient Segment

Visual salience is sometimes carelessly described as a physical property of a visual stimulus. It is important to remember that salience is the consequence of an interaction of a stimulus with other stimuli, as well as with a visual system (biological or artificial). As a straight-forward example, consider that a color-blind person will have a dramatically different experience of visual salience than a person with normal color vision, even when both look at exactly the same physical scene (see, e.g., the first example image below). As a more controversial example, it may be that expertise changes the salience of some stimuli for some observers. Nevertheless, because visual salience arises from fairly low-level and stereotypical computations in the early stages of visual processing the factors contributing to salience are generally quite comparable from one observer to the next, leading to similar experiences across a range of observers and of behavioural conditions.

6. Motion Saliency

The extraction of moving regions from sequential images is carried out by using BM. This kind of BM involves the loss of image information compared with the color BM using RGB and LAB color space models. Depicts the extracted result of moving regions by gray-scale BM, which shows the image information is excessively attenuated. LAB COLOR SPACE :

A *Lab* color space is a color-opponent space with dimension *L* for lightness and ***a*** and ***b*** for the color-opponent dimensions, based on nonlinearly compressed(e.g. CIE XYZ color space) coordinates.

The L*a*b* colour space includes all perceivable colors, which means that its gamut exceeds those of the RGB and CMYK color models (for example, RGB includes about 90% all perceivable colors). One of the most important attributes of the L*a*b*-model is device independence. This means that the colors are defined independent of their nature of creation or the device they are displayed on. The L*a*b* color space is used when graphics for print have to be converted from RGB to CMYK, as the L*a*b* gamut includes both the RGB and CMYK gamut. Also it is used as an interchange format between different devices as for its device independency. The space itself is a three-dimensional Real number space, that contains an infinite possible representations of colors. However, in practice, the space is usually mapped onto a three-dimensional integer space for device-independent digital representation, and for these reasons, the *L**, *a**, and *b** values are usually absolute, with a pre-defined range.

CIE L*a*b* (CIELAB*)* is the most complete color space specified by the International Commission on Illumination. It describes all the colors visible to the human eye and was created to serve as a device-independent model to be used as a reference.

The three coordinates of CIELAB represent the lightness of the color (L* = 0 yields black and L* = 100 indicates diffuse white; specular white may be higher), its position between red/magenta and green (a*, negative values indicate green while positive values indicate magenta) and its position between yellow and blue (b*, negative values indicate blue and positive values indicate yellow). The asterisk (*) after *L*, *a* and *b* are pronounced *star* and are part of the full name, since they represent *L**, *a** and *b**, to distinguish them from Hunter's *L*, *a*, and *b*, described below.

Since the *L*a*b** model is a three-dimensional model, it can be represented properly only in a three dimensional space. Two dimensional depictions include chromaticity diagrams: sections of the color solid with a fixed lightness. It is crucial to realize that the visual representations of the full gamut of colors in this model are never accurate; they are there just to help in understanding the concept.

Because the red-green and yellow-blue opponent channels are computed as differences of lightness transformations of (putative) cone responses, CIELAB is a chromatic value color space.

RGB color model is employed to prevent this excessive attenuation. Also, RGB color model has the shorter execution time because any additional image transformation is not required. But, it is a crucial disadvantage to be very sensitive to even small changes caused by light scattering or reflection. The parameter is proposed to overcome the sensitivity problem

$$\begin{bmatrix} M_i(x,y) \\ N_i(x,y) \end{bmatrix} \triangleq \begin{bmatrix} \min\left\{V_i^z(x,y)-\delta\right\} \\ \max\left\{V_i^z(x,y)+\delta\right\} \end{bmatrix},$$

$$i=\{r,g,b\},\ 0\le\delta\le 255$$

$$B_i^z(x,y)=\begin{cases} V_i^z(x,y), & \begin{cases} V_i^z(x,y)>M_i(x,y) \\ V_i^z(x,y)<N_i(x,y) \end{cases} \\ 0, \text{ else} \end{cases}$$

The moving regions extracted by are affected by the sensitivity parameter. To obtain the best image, this parameter can be adjusted according to the circumstances where the camera is installed. In our case, the best value is 18/2

The noise caused by light scattering or reflection can be eliminated by the proposed sensitivity parameter. However, the parameter should become larger to eliminate the noise caused by natural objects such as leaves and birds, which leads to extra attenuation on the moving regions. So the morphology, one of the geometric image processing schemes, is used to deal with this kind of noise appearing in the form of the crowd of pixels that the arrows indicate. The erosion operation of morphology removes the noises spread irregularly, and the dilation operation of morphology recovers the loss of moving regions made in the procedure of the erosion

Motion estimation is based on temporal changes in image intensities. The underlying supposition behind motion estimation is that the patterns corresponding to objects and background in a frame of video sequence move within the frame to form corresponding objects on the subsequent frame. Motion estimation is accomplished using ARP's algorithm.

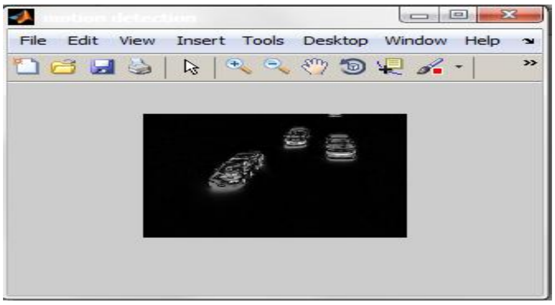


Figure 3: Motion estimated segment

After the determination of motion and visual salient features they are combined .combined images are detected then foreground objects are detected as shown below:

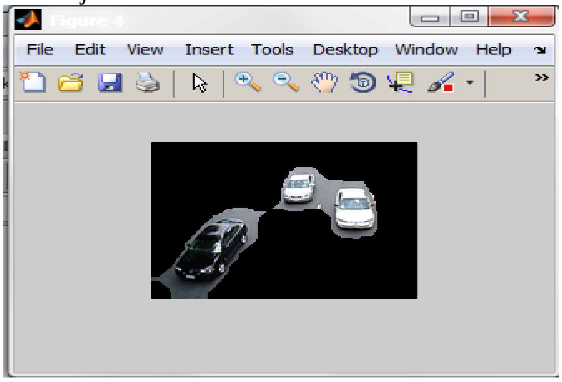


Figure 4: Foreground detected segment

7. Simulation Results

The parameters which determine the performance are PSNR AND MSE.

PSNR:it is the ratio of maximum of power signal to that of corrupted noise.for an image psnr can be calculated from mean square error.

MSE:Difference between estimator and estimated

$$MSE = \frac{1}{m\,n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j)-K(i,j)]^2$$

Where I and K represent the image and noise respectively and m,n represents row and column respectively.

$$\begin{aligned} PSNR &= 10\cdot\log_{10}\left(\frac{MAX_I^2}{MSE}\right) \\ &= 20\cdot\log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \\ &= 20\cdot\log_{10}(MAX_I) - 10\cdot\log_{10}(MSE) \end{aligned}$$

Comparing the SGMM and salient method we can conclude that better performance is shown by saliency detection .high PSNR value determines the quality of video detection .The graphical plot has been shown below:

Psnr 1 is representing the proposed method and psnr for the existing .It shows improvement leading to performance increase.

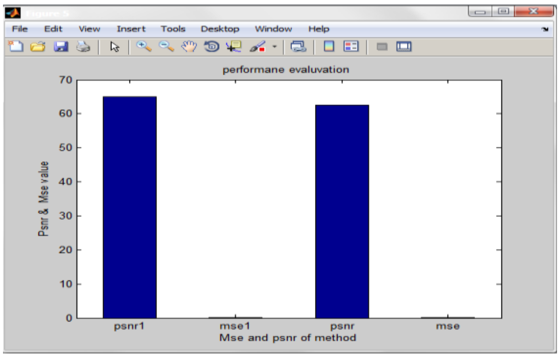


Figure 5: Performance evaluation of SGMM and saliency

8. Conclusion

We proposed a VOE approach which make use of motion and visual saliency induced features, such as shape, foreground/background colour models, and visual saliency, to extract the foreground objects in videos. We advanced a CRF model to integrate the above features, and additional constraints were introduced into our CRF model for preserving both spatial continuity and temporal consistency when performing VOE. Compared with SGMM this was shown better for the extraction of the foreground object due to the fusion of multiple types of saliency-induced features .high PSNR value is also obtained .

9. Acknowledgement

First of all we sincerely thank the almighty who is most beneficent and merciful for giving us knowledge and courage to complete the project work successfully. We also express our gratitude to all the teaching and non-teaching staff of the college especially to our department for their encouragement and help done during our work. Finally, we appreciate the patience and solid support of our parents and enthusiastic friends for their encouragement and moral support for this effort.

References

[1] R. Achanta and S. Süsstrunk, “Saliency detection using maximum symmetric surround,” inProc. IEEE Int. Conf. Image Process., Sep. 2010,pp. 2653–2656

[2] Y. Zhai and M. Shah, “Visual attention detection in video sequences using spatiotemporal cues,” inProc. ACM Int. Conf. Multimedia, 2006,pp. 815–824

[3] J. Zhong and S. Sclaroff, “Segmenting foreground objects from a dynamic textured background via a robust Kalman filter,” inProc. 9th IEEE Int. Conf. Comput. Vis., vol. 1. Oct. 2003, pp. 44–50.

[4] F. Liu and M. Gleicher, “Learning color and locality cues for moving object detection and segmentation,” inProc. IEEE Conf. Comput. Vis.Pattern Recognit., Jun. 2009, pp. 320–327

[5] M. Grundmann, V. Kwatra, M. Han, and I.Essa, “Efficient hierarchical graph based video segmentation,” in Proc. IEEE Conf. Comput. Vis.Pattern Recognit., Jun. 2010, pp. 2141–2148

[6] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,”IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645,Sep. 2010

References

Aiswarya Muralidharan did her Bachelor of Technology in Electronics and communication at Amaljyothi college of Engineering, kottayam, Kerala and pursuing her Master of Engineering in Communication system at Sri Shakthi Institute of Engineering and Technology,Coimbatore,Tamilnadu. She has presented one paper in International Conference and one paper in National Conference and published a journal.

S.Sivakumar completed his Master of Engineering in Communication System and presently working as Assistant professor in Sri Shakthi Institute of Engineering and Technology, Coimbatore and has four and half years of experience His area of interest is Antenna design. He published three journals