Object Tracking In Real Time Video Applications Using Semantic Segmentation

P.Saravanarathinam[1], Dr.K.Ramar[2]

[1]*M.E. Scholar, Dept. Of CSE, Einstein college of Engineering, Tirunelveli, saro433@gmail.com*

[2]*Principal, Einstein College of Engineering, Tirunelveli.*

***Abstract*— Video tracking is the process of locating a moving object (or multiple objects) over a time. The objective of video tracking is to associate target objects in consecutive video frames. The association can be especially difficult when the objects are moving fast relative to the frame rate. The proposed system is designed to achieve the video tracking using semantic segmentation. This system is proposed to find the over speed vehicle from the video samples with efficient algorithms and approaches. First Initial segmentation is done based on the colour as a segment parameter. Semi-Semantic segmentation is performed to find required objects appropriately. Finally Semantic segmentation is done to point the car objects in the given video input.The speed of the detected car objects is estimated based on its time and distance travelled by those objects throughout the video. Based on the threshold value, the over speed vehicle is detected. The high speed vehicle is highlighted in the output video.**

Keywords*- vehicle detection, *semantic segmentation, geodesic propagation, video tracking

I. INTRODUCTION

Image Segmentation is the process of partitioning a digital image into multiple segments .The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics.

The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image. Each of the pixels in a region are similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different with respect to the same characteristic(s). When applied to a stack of images, typical in medical imaging, the resulting contours after image segmentation can be used to create 3D reconstructions with the help of interpolation algorithms like Marching cubes.

Semantic Labeling, sometimes also called shallow semantic parsing, is a task in natural language processing consisting of the detection of the semantic arguments associated with the predicate or verb of a sentence and their classification into their specific roles.

Feature Extraction starts from an initial set of measured data and builds derived values intended to be informative, non redundant, facilitating the subsequent learning and generalization steps, in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be redundant, then it can be transformed into a reduced set of features. This process is called feature extraction. The extracted features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data. Feature extraction involves reducing the amount of resources required to describe a large set of data. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which over fits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

Video Tracking is the process of locating a moving object (or multiple objects) over time using a camera. It has a variety of uses, some of which are: human-computer interaction, security and surveillance, video communication and compression, augmented reality, traffic control, medical imaging and video editing. Video tracking can be a time consuming process due to the amount of data that is contained in video. Adding further to the complexity is the possible need to use object recognition techniques for tracking, a challenging problem in its own right.

The objective of video tracking is to associate target objects in consecutive video frames. The association can be especially difficult when the objects are moving fast relative to the frame rate. Another situation that increases the complexity of the problem is when the tracked object changes

orientation over time. For these situations video tracking systems usually employ a motion model which describes how the image of the target might change for different possible motions of the object.

Examples of simple motion models are:

- When tracking planar objects, the motion model is a 2D transformation (affine transformation or homography) of an image of the object (e.g. the initial frame).
- When the target is a rigid 3D object, the motion model defines its aspect depending on its 3D position and orientation.
- For video compression, key frames are divided into macro blocks. The motion model is a disruption of a key frame, where each macro block is translated by a motion vector given by the motion parameters.
- The image of deformable objects can be covered with a mesh, the motion of the object is defined by the position of the nodes of the mesh.

To perform video tracking an algorithm analyzes sequential video frames and outputs the movement of targets between the frames. There are a variety of algorithms, each having strengths and weaknesses. Considering the intended use is important when choosing which algorithm to use. There are two major components of a visual tracking system: target representation and localization, as well as filtering and data association.

Target representation and localization is mostly a bottom-up process. These methods give a variety of tools for identifying the moving object. Locating and tracking the target object successfully is dependent on the algorithm. For example, using blob tracking is useful for identifying human movement because a person's profile changes dynamically. Typically the computational complexity for these algorithms is low.

II. OUTLINE OF PROPOSED SYSTEM

A. OVER ALL SYSTEM DESIGN

This work uses the geodesic propagation based semantic image segmentation method.First the input video is selected and the splitting of frames process is done for further process. The initial segmentation is performed to segment the image based on the colour value using four connected graph. Feature extraction is done by using the GIST algorithm and assigns the labels for the extracted features .Semantic labelling is done to find out the remaining parts. The required objects is extracted in the semantic segmentation phase by using the geodesic propagation approach. Finally speed estimation is done to calculate the speed of the vehicle with the help of the position of the vehicle in each frame. Based on the threshold value, the object is marked as high speed vehicle.

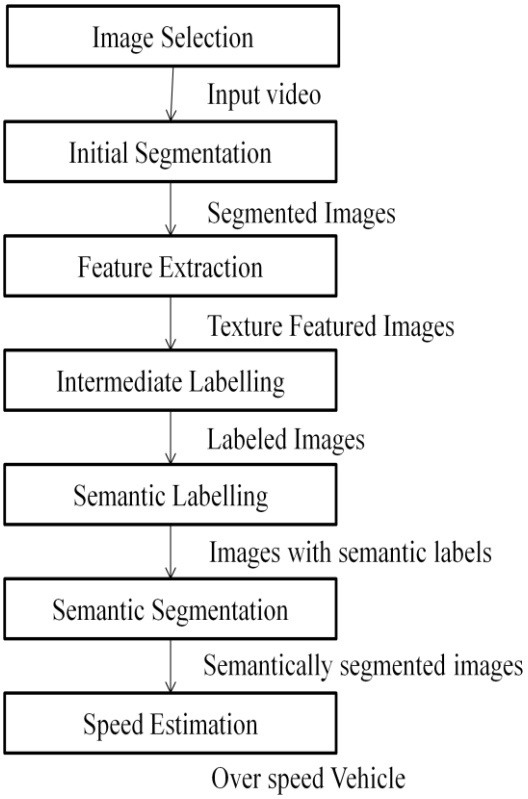

Fig 1: Over all structure of this system

III. METHODOLOGY

The proposed system is used to find the over speed vehicle using semantic segmentation and geodesic approach. The proposed system consists of many process to extract the required object. Initially, the **Semi-Semantic Segmentation** is done. In this task Initial segmentation is done to segment the frames of images using colour values using 4-connected graph. Feature Extraction is performed to extract the important features in the initially segmented frames using the GIST algorithm Assign intermediate labelling to the extracted features for the identification purpose .**Semantic labelling** is performed to find out parts of the objects which is left out. **Semantic Segmentation** is performed using geodesic approach to extract the exact car object from the frames. **Speed Estimation** is to calculate the overall speed of the vehicle by using the distance and time travelled .Finally the vehicle is determined as over speed vehicle if its speed is above the threshold value. Then mark the car as over speed vehicle and the output video with marked object is produced.

Image Selection

This is the initial process. In this process the input image is selected for doing the segmentation. The input image is selected using the browser control. The browse control gets pathname and the file name of the image. And then the input image is converted into the matrix. And then the selected input image is displayed in the axes control. The

image may be any format jpg, png, bmp etc. But the input image must be color.

Assign Semantic Label

After getting the input image the next process is to assign the semantic label. These labels must be assigned to each pixel in the input image. To assign the semantic label, the 4 connected graph method is used. This method finds the distance between the adjacent (neighbor) pixels of each pixel. These difference values consider as the distance between two pixels and then assign this value as semantic label for that pixel. Finally the labeled image is produced for further process.

Feature Extraction

The third process is the Feature extraction. It means to retrieve the important features (texture information) from the input image. To retrieve the texture features the GIST feature extraction technique is used. GIST feature extraction technique uses the gabor filter with different scale and orientation. Retrieval based on Gabor features. The following algorithm is used to extract the feature.

1. First get the input image.
2. Generate the Gabor filter with different scale and orientation
3. Convolve the image with 17 Gabor filters and to produce 17 feature maps.
4. Concatenate the averaged values of all 17 feature maps to produce the final feature value. This final feature is called as GIST features.

Generation of proposal map

After extracting GIST features the next step is to generate the proposal map. To generate the proposal map first find the GIST features for all images in the training dataset by using GIST retrieval algorithm.

Then get the extracted GIST feature values for the input image and all the images in the training dataset. Then these features are given into the Join adaboost algorithm for recognition object in the given input image. The adaboost algorithm first gets the feature values of the all training images. From the training feature values all the object models are trained and then input image feature values classified based on the trained object model based on the weak classifier. Finally the recognized object name is produced as the result value. Using this value to retrieve all ground truth images corresponding to the recognized object name. And then find the difference value between the retrieved images from the training dataset and the input image for each pixel. Then the minimum difference value and its corresponding image is chosen as the closest image of the given input image. So the ground truth image of the chosen image from the training dataset as used as the proposal map of the input image. Then find the different color value from the produced proposal map. These count value are considered as the number of objects in the given input image. These proposal map is not considered as the efficient map.

So, to produce the efficient proposal map, the difference of the cluster center between the input image and the retrieved images from the training data set is used. To find the cluster center for retrieved images and original images, the FCM (Fuzzy C-Means) segmentation algorithm is used. The algorithm is as follows.

Let $X = \{x_1, x_2, x_3 ..., x_n\}$ be the set of data points and $V = \{v_1, v_2, v_3 ..., v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the fuzzy membership $'\mu_{ij}'$ using:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

3) Compute the fuzzy centers $'v_j'$ using:

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...c$$

4) Repeat step (2) and (3) until the minimum $'J'$ value is achieved or $||U^{(k+1)} - U^{(k)}|| < \beta$.

where,

'k' is the iteration step.

$'\beta'$ is the termination criterion between [0, 1]

$'U = (\mu_{ij})_{n*c}'$ is the fuzzy membership matrix.

'J' is the objective function.

After FCM Segmentation cluster centers values are produced for both the input images and the retrieved image from the training dataset. Then find the difference between two cluster centers (input image and training images). Then the minimum difference value and its corresponding image is chosen as the closest image of the given input image. So the ground truth image of the chosen image from the training dataset as used as the proposal map of the input image. Finally the proposal map of the input image is produced

.

Seed Localization

The fifth process of the Geodesic Segmentation is the Seed Localization. To locate the seed the proposal map of the input image is used. To find the initial seeds apply the mean shift segmentation. The mean shift algorithm as follows.

Mean shift Segmentation Algorithm

1. Consider the given image.
2. The image is divided into pixels.
3. n be the total number of pixels.
4. Find the feature vector set and initialized
5. The mean shift segmentation is defined as

$$m_G(y_i) = \frac{\sum_{j=1}^{n} x_j g(\| \frac{y_i - x_j}{h} \|^2)}{\sum_{j=1}^{n} g(\| \frac{y_i - x_j}{h} \|^2)} - y_i$$

After the segmentation of the proposal map, select 10 pixels are randomly from the each group of the segmented image as seed values of the each group. These seed values are considered as the initial seeds.

Edge weight calculation

After finding the seed value the next process is to find the Edge weight. To find the edge weight the EM Segmentation algorithm is used. The following formula is used for weight calculation.

$$w^c(x, x'|l) = \frac{\|p(l|x) - p(l|x')\|}{p(l|x) + p(l|x')}.$$

Geodesic Propagation

The Final process is the Geodesic Propagation. This process is used to segment the input image based on the object. Geodesic Propagation find the geodesic distance of all object classes efficiently. For a pixel x, if one label li is propagated to x earlier than other labels, then the corresponding geodesic distance Dl (x) to li is shorter than others. It propagates all labels simultaneously to the entire image, and once the geodesic path of label li reaches pixel x, its shortest geodesic distance minl∈LDl (x) is determined.

During geodesic propagation, each vertex has three statuses: labelled, reachable and unlabeled. The labelled vertex is assigned label determinately, as well as its minimal geodesic distance. The set of reachable vertices includes the neighbours around the labelled vertices. The reachable vertices are sorted according to their geodesic distance and put into the ordered queue QR. Other vertices are marked as unlabeled to indicate that the geodesic propagation has not reached them yet. iteratively selects the vertex vi of the minimum distance in the reachable queue QR, sets vi as labelled, and propagates labels to its neighbouring vertices, until the reachable queue is empty.

IV. EXISTING TECHNOLOGY

A Kalman filter is an optimal estimator. It infers parameters of interest from indirect, inaccurate and uncertain observations. PSO based algorithm is used for multi-target tracking. At the beginning of each frame the targets are tracked individually using highly discriminative appearance models among different targets. Particle filters are based on probabilistic representations of states by a set of samples. Most of the existing approaches utilize and integrate low level local features and high-level contextual cues. However, the lack of meaning in the primitives and the cues provides low discriminatory capabilities, since they are rarely object-consistent. Moreover, blind combinations of heterogeneous features and contextual cues exploitation through limited neighbourhood relations in the CRFs tend to degrade the labelling performance. But there are difficulties for it to overcome problems like object deformation, especially when the multiple objects move too close. Kalman Filter misunderstood and considered those multiple objects as only one object. PSO methods required more time. If the target object did not reappear in close region, typical PSO methods were hardly retrieve the target because low performance of particle diversity. The drawback of Particle Filter in its application. For fast moving object and motion blur, Particle Filter plays inefficient role due to loss of particle diversity and repeat the selection and distribution when reach a local optima Object Segmentation method takes longer computational time.It takes high memory cost.If the target object did not reappear in close region, typical segmentation methods were hardly retrieve the target because low performance of particle diversity.

V. CONCLUSIONS

This work uses the geodesic propagation based semantic image segmentation method. First the input image is given into the semantic segmentation. Semantic image segmentation is a fundamental yet challenging problem. Then the semi-semantic segmentation method is used to effectively detects object parts. And then the technique is introduced to transform the visual space into a higher level space with bootstrap and fuzzy c-means segmentation values as intermediate labels. Finally, geodesic based segmentation is introduced for the final semantic labelling.

In future instead of fuzzy c-means segmentation other segmentation such as mean shift, k-means will be used. Not only that instead of geodesic segmentation approach other highly efficient approach will be consider for improving efficiency.

REFERENCES

[1] M. Valera and S. Velastin, ―Intelligent distributed surveillance systems: A review,‖ IEE Proc. Vis. Image, Signal Process., vol. 152, no. 2, pp. 192–204, Apr. 2005.

[2] J.-W. Hsieh, Y.-T. Hsu, H.-Y. M. Liao, and C.-C. Chen, ―Video based human movement analysis and its Application to surveillance systems,‖ IEEE Trans. Multimedia, vol. 10, no. 3, pp. 372–384, Apr. 2008.

[3] D. A. Migliore, M. Matteucci, and M. Naccari, ―View-based detection and analysis of periodic motion,‖ in Proc. 4th ACM Int. Workshop Video Surveill. Sensor Netw., Oct. 2006, pp. 215–218.

[4] S.-Y. Chien, S.-Y. Ma, and L.-G. Chen, ―Efficient moving object segmentation algorithm using background registration technique,‖ IEEE Trans. Circuits Syst. Video Technol., vol. 12, no. 7, pp. 577–586, Dec. 2002.

[5] W.-K. Chan and S.-Y. Chien, ―Real-time memory-efficient video object segmentation in dynamic background with multi-background registration technique,‖ in Proc. IEEE Multimedia Signal Process. Workshop, Oct. 2007, pp. 219–222.

[6] S.-Y. Chien, Y.-W. Huang, B.-Y. Hsieh, and L.-G. Chen, ―Single chip video segmentation system with a programmable PE array,‖ in Proc. IEEE Asia-Pacific Conf., Aug. 2002, pp. 233–236.

[7] W.-K. Chan, J.-Y. Chang, T.-W. Chen, and S.-Y. Chien, ―Efficient content analysis engine for visual surveillance network,‖ IEEE Trans. Circuits Syst. Video Technol., vol. 19, no. 5, pp. 693–703, May 2009.

[8] D.-Z. Peng, C.-Y. Lin, W.-T. Sheu, and T.-H. Tsai, ―A low cost and low complexity foreground object segmentation architecture design with multi-model background maintenance algorithm,‖ in Proc. IEEE Int. Conf. Image Process., Nov. 2009, pp. 3241–3244.

[9] Mittal and N. Paragios, ―Motion-based background subtraction using adaptive kernel density estimation,‖ in Proc. IEEE Conf. Comput. Vision Patt. Recog., Jun.– Jul. 2004, pp. 302–309.

[10] Abhijit Kundu, Yin Li, Frank Dellaert, Fuxin Li, and James M. Rehg, ―Joint Semantic Segmentation and 3D Reconstruction from Monocular Video‖