Vehicle Tracking using Convolutional Neural Network

S. Shruthi, *Member, IAENG*

Abstract: **In this paper, vehicle tracking is done as a learning problem of estimating the location and the scaling of an object in a nearby location, given its previous location. This scales the current and previous image frames. Convolutional Neural Networks (CNNs) are trained to perform the above estimation task. Using CNN we study the dynamic state features. The changes of the target's nearby visual appearance often lead to tracking failure in practice which is alleviated by Improved Shift-variant CNN architecture. We present a novel tracking method for effectively tracking vehicles in structured environment where state subtraction of the object (vehicle) is extracted and motion tracking is considered as an important criterion. An efficient and robust method has been implemented in this paper which is called as DTT (Detection, Tracking and Tracker). Here, segmentation is considered as the detection process, motion tracking is considered as the tracking process and improved shift-variant is considered as the tracker. This kind of tracking is 'object independent' and so the proposed method can be applied to track other objects too.**

Index Terms: **Convolutional Neural Networks, Color Segmentation, Motion tracking, vehicle tracking.**

I.INTRODUCTION

The most fundamental problem in computer vision is object tracking. This is a relatively easy task when the objects are isolated and easily distinguished from the background. However, in complex and crowded environments, many objects are present that may have similar appearances, and occlude one another; also occlusions by other scene objects are common. The proposed method can robustly track multiple objects (vehicle) under such challenging conditions. Tracking Traditional feature-based methods, such as those based on color [1] or motion blobs [2]–[4], perform tracking by maintaining a simple model of the target and adapting such a model over time.

The major challenge of the traditional learning-based and or tracking-by-detection methods is the false positive matches that lead to wrong association of the tracks. The reason is that those methods are based on applying an appearance model or object detector at all possible windows around the target, and the object detection in the current frame.

When there is a heavy traffic and when the distracting objects are similar to the target, the object detector will generate similar high detection scores for both the target and the distracters, which might probably cause a drift problem. This kind of drift problem is alleviated using shift-variant CNN architecture.

An important stage in automatic vehicle crash monitoring systems is the detection of vehicles in each video frame and accurately tracking the vehicles across multiple frames. Given the detected vehicles, tracking can be viewed as a correspondence problem in which the goal is to determine which detected vehicle in the next frame corresponds to a given vehicle in the current frame.

In this paper, we use Convolutional Neural Networks (CNNs) [10] as our base learner because they have been demonstrated to be able to extract local visual features (structures) and they are widely used in various visual recognition applications [11], [12]. Different from fully connected neural networks, CNNs force the extraction of local features by restricting the receptive fields of hidden units to be local, based on the fact that images have strong 2-D local structures. In this paper, we use Convolutional Neural Networks (CNNs) [10] as our base learner because they have been demonstrated to be able to extract local visual features (structures) and they are widely used in various visual recognition applications [11], [12]. Different from fully connected neural networks, CNNs force the extraction of local features by restricting the receptive fields of hidden units to be local, based on the fact that images have strong 2-D local structures. From the conventional shift-invariant CNN we move to improved shift-variant architecture. Shift-invariant CNN (see Section III (c) or [13]) can make them suitable for recognition or detection tasks but inappropriate for tracking tasks.

The improved shift-variant CNN is designed as a CNN tracker. This plays a key role in turning the CNN model from a detector into a tracker object segmentation method. The features (structures) are learned during offline training. The features (structures) are learned during offline training. A contour associated to a moving region is initialized using a motion segmentation step which is based on image differencing between an acquired image and a continuously updated background image. The dynamic state processes are learned using motion tracking where distance transform and particle filtering are integrated. Dynamic states are modeled by the distance field using the distance transform and integrated into particle filtering for effective object tracking for surveillance videos. Once after dynamic state process has been extracted we go for shift-variant approach in-order to alleviate the drift problem and to track the adjacent image of vehicles.

Manuscript received March 06, 2011; revised April 12, 2011

S. Shruthi is with the Panimalar engineering college, Chennai 602103, Tamil Nadu, India. She is a student in the Department of Computer Science Engineering (Phone: +91-967-713-0645; e-mail: shruthigeethu@gmail.com).

The contributions of this paper include the following: 1) a discriminative model extracts dynamic state features for specific object class tracking, where the features are learned from a parametric feature pool with rich degrees of freedom; 2) the shift-variant architecture of which is a alternative to traditional use of CNN for alleviating drift problem; and 3) color segmentation and state subtraction is acts as a key idea for detecting the objects.

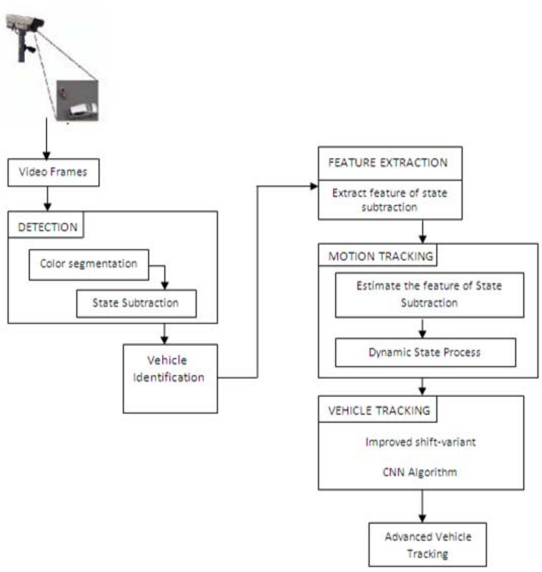


Fig: 1. Block diagram of the proposed Vehicle Tracking System

The rest of this paper is organized as follows. Section II describes color segmentation and state subtraction. Sections III discuss the details of the CNN motion tracking algorithm and shift-variant CNN architecture. Section IV summarizes the conclusion.

II.Colour Segmentation And State Subtraction

The simplest technique for separating moving objects from a stationary background is through examining the difference between each new frame and an estimate of the stationary state.

Segmentation of moving objects in an outdoor environment requires in addition that the background estimate evolve over time as lighting conditions change. Changes due to moving objects must be differentiated from the changes in stationary parts of the image. Here, a modified version of the moving is suggested by [Karmann & Von Brandt 90] and [Kilger 92]. This method uses an adaptive state model. The state model is updated in kalman filter formalism, thus allowing for dynamics in the model as lighting conditions change. The state is updated each frame via the update equation.

$$St+1= St + (\alpha1 (1-Mt) + \alpha2Mt) Dt \qquad (1)$$

Where St Represents the state model at time t, Dt is the difference between the present frame and the state model, and Mt is the binary moving objects hypothesis mask.

The gains α1 and α2 are based on an estimate of the rate of change of the state. In a complete Kalman filter implementation, these values would be estimated along with the state since they correspond to the elements of the error covariance. The hypothesis mask, Mt, attempts to identify moving objects in the current frame.

The state subtraction is achieved by subtracting the state from the current input frame to detect the vehicles. Here, frame at time t from the input video along with the previously acquired background frame (where no vehicles are found) is fed as input to the algorithm. The algorithm subtracts the intensity value of each pixel in the input frame

It(x,y) from the state image Ibk(x,y) resulting in a difference image Idiff(x,y) given by,

$$Idiff(x,y) = [It(x,y) - Ibk(x,y)] \qquad (2)$$

The state subtraction step is performed to detect moving objects since the static objects are part of state. Thus, we are left with the intensity values of moving objects in the difference image Idiff(x,y).

III.Convolutional Neural Network

A. Motion Tracking

Here temporal features are considered to track the motion of a vehicle. Dynamic state process is implemented using motion tracking. Dynamic background may not be available when considering the temporal and spatial features but this is overcome by motion tracking algorithm. Filtering and data association provides a flexible framework for modeling objects (vehicle) and the environments as the environment state into the state vector of the object (vehicle). The environment state is modeled by the distance between the object (vehicle) and environment boundaries, because the state reflects the motion pattern of the object (vehicle). When objects move parallel to the boundary, such as people walking along the lane, such distances are increasing/decreasing gradually. This problem is solved by combining the distance transform and particle filtering.

In vehicle tracking, the state vector only includes the dynamic characteristics of the object, e.g., location, orientation, scale, etc. Here dynamic state is denoted by **y**. In structured environment, object motion is constrained by the environment structure, and the relationship between the object of interest and the environment is also an important characteristic for estimating the object motion. Environment state, models the environment related characteristics of the object, and is denoted by **e**. The environment state may include, but is not limited to, the geometric relation between the object and environment entities, the property of the region which the object (vehicle) is currently in, etc. The state vector at time **t, xt**, for the object within adaptive state subtraction method is employed to extract a large set of objet motion patterns from the videos over a period of time. Hu et al. [5] proposed a method for predicting motion and detecting the abnormal activities from videos which is based on the learning of statistical motion patterns. Considering that the motions of objects are constrained by the environment, we explore the relationship between the objects (vehicle) are constrained by the environment, relationship between the objects (vehicle) and the environments are explored as the high-level information to help tracking the objects (vehicle). Tracking the environments in the framework of motion tracking by incorporating the relationship between the object problems in the structured environments is comprised of two sub states:

1) The dynamic state, yt;
2) The environment state, et.

The current environment state depends upon previous environment states. Current dynamics and environment states are coupled by the distance map.

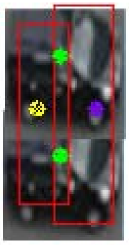
B. CNN

Visual recognition is best performed by CNN. For visual recognition tasks, CNNs first extract and combine local features from the input image and these features are then combined by the subsequent layers in the order to obtain higher order features. Such high order features are eventually encoded into 1-D vector (target label), which is then categorized by a trainable classifier. It is worth nothing that features extraction is a non-trivial problem, because there are size, slant, and position variations, CNN combines 2 process: Tracking (Dynamic state process) and, tracker (Shift-variant architecture). The basic idea is that, given two corresponding image patches where in the first patch the target is located at the center, we want to find the target location in the second image patch on the basis of motion of the object(vehicle) tracking (spatial and temporal structures) of these two patches.

The dynamic state process indicates the appearances while the temporal structures capture the motion information. Here state subtraction is made reliable and it is also available for dynamic background. For tracking problem, assume that the target position at the time t-1 is known, the goal is to find the target position at time t. Denote the target position at time t-1 by $x_{t-1}=(x_{t-1}\ ,t_{t-1}\ ,s_{t-1})$, where $(x_{t-1}\ ,y_{t-1})$ is the position of the vehicle, s_{t-1} is the scale. The associated bounding box is denoted by R (x_{t-1}). We extract an image patch which includes the surrounding of the target at time t-1. Denote the image patch by $N_{t-1}(x_{t-1})$. The ratio of the size of $N_{t-1}(x_{t-1})$ to that of R (x_{t-1}) is fixed (the entire vehicle can be included in $N_{t-1}(x_{t-1})$ for most cases, so ratio is set. We extract the image patch $N_t(x_{t-1})$ at time t which has the same position as $N_{t-1}(x_{t-1})$. $N_{t-1}(x_{t-1})$ and $N_t(x_{t-1})$ are normalized to image patches of fixed size w x h which are the inputs of CNN. After normalization, the target object is always at the same position O in the w x h input patch at time t-1. This CNN is expected to detect an object at time t which corresponds to the target located at position O. The output of CNN is the probability map of size (w/2) x (h/2) which shows the target position at time t.

The advantages of using probability map is that, 1) only around the target center the probability of the unit is high, so the probability map reflects an accurate localization property, and 2) we can measure how well the probability map describes the true situation, i.e., the nosier the probability map, the more complex is the environment and therefore, the less confident we are in using the map to estimate the target position.

C. SHIFT-VARIANT CNN



Green denotes-O
Yellow denotes-$H_{V,}$ H_V'
Purple denotes-K_V, K_V'

Fig: 2. Improved shift-variant approach

The major difference between shift-variant CNN and shift-invariant CNN is explained clearly. With the help of Improved shift-variant adjacent image of vehicles are tracked efficiently. Modeling objects by bounding boxes which is considered to be inappropriate for more complex cases, e.g., vehicles not view from near top-view angles, more sophisticated modeling of objects is overcome by this paper. But this is not possible in 3-D space. The significance of shift-variant is explained below.

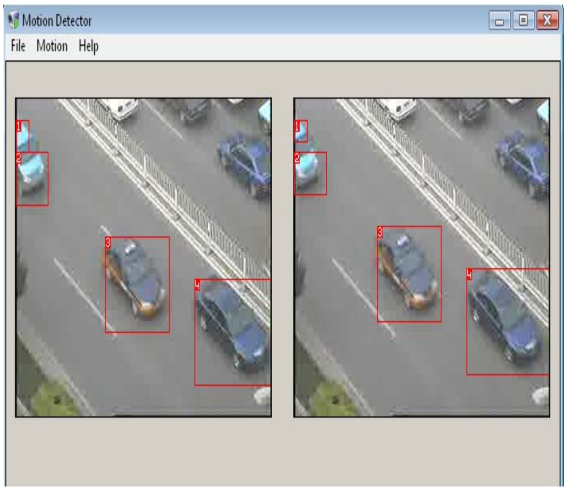


Fig: 2.0. Improved Shift-Variant CNN-difference between the different data-sets are extracted

Denote the target position at time **t-1** and time t by O and **V** respectively. Consider there is another vehicle moving from O' to V' at the same time. The global and local receptive field of V is denoted by H_V and $K_{V,}$ respectively. Denote the center of the receptive field H_V by C_{Hv}. The probability p_V in the probability map is the sigmoid function of the summation of H_A and K_V responses, which is monotonically increasing.

The conventional architecture of CNNs for detection is the so-called space displacement neural network (SDNN) [13], [14]. In this improved shift-variant CNN, a single detector is replicated over the input, and the output is a detection score map which is shift-invariant. This means that the position *V* and *V’* will generate similar high detection scores. The high detection score at *V*' is very harmful because it may cause confusion with *V* (the readers can refer to Fig. 1 for the intuition). If we use two-times up sampling from layer *C*3 to the output (similar to the architecture in [14]), our network will be a special instance of SDNN, which operates as an object detector. Consider C_{Hv} is at V. It means that the relative locations of the global receptive field (*CGA*) and the object position (*A*) is the same. Hence the obtained detection map is shift-invariant. The use of four-times up sampling has a nice property. The center of the global receptive field of one output unit is at the midpoint of the previous location and the current location of the object (Property 3.1, the proof is shown in Appendix). Having this property, the four-times up sampling breaks the shift-invariant property of SDNN. In the Improved shift-variant CNN, C_{Hv} depends on O so the CNN output depends on the Object’s previous location. This is different from the shift-invariant CNN where is independent of *O* (it is at *A*). Hence, the proposed model is different from conventional CNN models which operate as object detectors. In this Improved shift-variant, we claim that the tracker has less chance to drift to *V’* by showing that *pA* is larger than *pA_*. As *pA* is the sigmoid function of the summation of H_V and $K_{V,}$ responses, it is large only if both H_V and $K_{V,}$ respond large. For *V* and *V’*, the response of the local branch K_V and K_V’, are similar, as they operate as local object detectors. For the response of the global branch and H_V’, H_V responds large, as it extracts the similar pattern at *O* (at time *t*−1) and *V* (at time *t*) (this is learned during training).

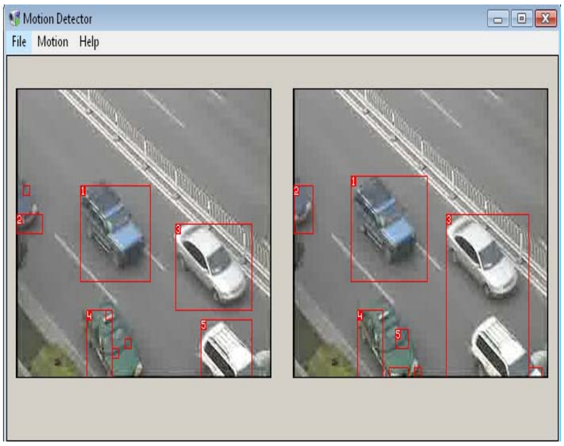


Fig: 2.1. a) Improved Shift-Variant b) Shift-Variant

But H_V’ responds small, because the global patterns at *O* (at time *t* −1) and *V’* (at time *t*) are different. Therefore, *pA* is larger than *pA_*, and the tracker does not drift to *V’*. From this point of view, the drift problem is alleviated. It is clear that the tracker does not drift to other locations, because the responses of the global and local receptive fields are both small and the top view angle of the object (vehicle) is captured easily.

IV.CONCLUSION

In this paper, tracking is done based on CNNs. Here Vehicle tracking is done by the proposed DTT method which proved to be efficient and robust. Color segmentation and state subtraction is done as a part of detection where dynamic state process is to improve the part of tracking. The Improved shift-variant extended the use of conventional CNNs and is considered to be a efficient tracker. Implementation of the above studies in 3-D will be my future study and CNN model is not designed to handle full and long-term occlusions by the distracter of the same object class.

REFERENCES

[1] Jialue Fan, Wei Xu, Ying Wu and Yihong Gong,” Human tracking using Convolutional Neural Network,” in Proc. IEEE transactions on neural networks, vol.21, NO.10, Oct 2010

[2] Junda Zhu, Yuanwei Lao, and Yuan F. Zheng, “Object tracking in structured environment for video surveillance applications”, IEEE transactions on circuits and systems for video technology, vol.20, February 2010.

[3] D. Koller, J. Weber and J. Malik, “Robust multiple car tracking with occlusion reasoning,” Proc. Third European Conference on Computer Vision, 1994, pp. 189-196, May 2-6 1994

[4] Logesh Vasu and Damon M. Chandler, “Vehicle tracking using a Human-Vision-based Model of Visual Similarity” IEEE 2010.

[5] R. T. Collins and Y. Liu, “On-line selection of discriminative tracking features,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Madison, WI, Jun. 2003, pp. 346–352

[6] S. Kamijo, Y. Matsushita, K. Ikeuchi and M. Sakauchi,“Traffic monitoring and accident detection at intersections,” IEEE Trans. On Intelligent Trans. System, vol 1, no 2, pp 108-118, June 2000

[7] C. Huang, B. Wu, and R. Nevatia, “Robust object tracking by hierarchical association of detection responses,” in Proc. Eur. Conf. Comput. Vis., Marseille, France, Oct. 2008, pp. 788–801.

[8] Giles, A. C. Tsoi, and A. D. Back, “Face recognition: A Convolutional Neural Network approach,” IEEE Trans. Neural Netw., vol. 8, no. 1, pp. 98–113, Jan. 1997.

[9] X. Song and R. Nevatia, “Detection and tracking of moving vehicles in crowded scenes,” IEEE Workshop on Motion and Video Computing, vol., no., pp. 4-4, February 2007.

[10] N.K. Kanhere, S.J. Pundik and S.T. Birchfield, “Vehicle segmentation and tracking from a low-angle axis camera,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, no., pp. 1152-1157, June 20-25 2005.