Vision-based Bicycle Detection and Tracking using a Deformable Part Model and an EKF Algorithm

Hyunggi Cho, Paul E. Rybski and Wende Zhang

***Abstract*— Bicycles that share the road with intelligent vehicles present particular challenges for automated perception systems. Bicycle detection is important because bicycles share the road with vehicles and can move at comparable speeds in urban environments. From a computer vision standpoint, bicycle detection is challenging as bicycle's appearance can change dramatically between viewpoints and a person riding on the bicycle is a non-rigid object. In this paper, we present a vision-based framework to detect and track bicycles that takes into account these issues. A mixture model of multiple viewpoints is defined and trained via a Support Vector Machine (SVM) to detect bicycles under a variety of circumstances. Each component of the model uses a part-based representation and known geometric context is used to improve overall detection efficiency. An extended Kalman filter (EKF) is used to estimate the position and velocity of the bicycle in vehicle coordinates. We demonstrate the effectiveness of this approach through a series of experiments run on video data of moving bicycles captured from a vehicle-mounted camera.**

I. INTRODUCTION

The automotive industry is increasingly interested in adding more intelligence to cars and trucks with the ultimate goal of developing fully autonomous automobile traffic. To this end one of the most important research areas to address is that of automated perception systems that will allow the vehicle to comprehend its immediate environment and make decisions that enhance the safety of vehicle occupants [11] as well as the safety of persons around it. This is especially true for the class of objects called vulnerable road users (VRUs) [8] which includes entities such as bicyclists, motorcyclists, pedestrians, and operators of other small vehicles as shown in Figure 1. A perception system that can, in real time, gather enough information to do a complete scene analysis is currently beyond the immediate scope of this work. Rather, we focus on the problem of identifying and extracting specific quantities of interest from the scene. In particular, we are interested in focusing on the problem of detecting and tracking bicyclists from an on-board vision system. In general, bicyclists and pedestrians are the most vulnerable of the class of VRUs due to the lack of any real protection against collisions. However, bicyclists move at speed equivalent to a slow moving vehicle and, by law, must share the road with vehicles in most urban environments. This puts them at particular risk for suffering life-threatening accidents.

Fig. 1. Of all the entities in the class of vulnerable road users (VRUs), pedestrians and bicyclists are the most likely to suffer severe injuries and death if they are involved in a collision with an automobile.

A number of researchers working on Intelligent Transportation Systems have proposed a number of different approaches for vehicle perception systems. Approaches using sensors such as vision, LIDAR, and RADAR have all been proposed as well as a number of systems which use the fusion of two or more of these sensors in order to provide more robust detection results [12] (see Section II for details on more approaches). In this work, we examine the use of video systems to detect and track bicyclists. Imagery from video cameras contains a wealth of high resolution information about the environment that can effectively be used to solve a number of perception problems. One of the most compelling arguments for using monocular cameras in automotive applications is that they are very inexpensive when compared to current LIDAR or RADAR systems. However, one of the biggest challenges when using computer vision systems to detect objects is handling the variations in the object's appearance, shape, and motion. Real-world environments can be very complex so that separating foreground from background is also a difficult problem. Finally, the motion of the vehicle that carries the camera must also be taken into account. Because pedestrians are ubiquitous, most research has focused primarily on them [10], and there is a comparative dearth of research in the detection and tracking of bicycles. While the two problems are similar in some ways, we believe that the bicycle problem is more challenging. For instance, the appearance of bicycles to a camera can change dramatically depending on the viewing angles. Additionally, the speed of bicyclists is much higher and their proximity to vehicles is typically much closer.

In this paper, we build a three-component bicycle model using *Felzenszwalb*'s deformable part-based model [7]. Then,

H. Cho and P. E. Rybski are with the Robotics Institute, Carnegie Mellon University, 5000, Forbes Ave., Pittsburgh, PA 15213, USA. {hyunggic, prybski}@cs.cmu.edu

W. Zhang is with the Electrical and Controls Integration Lab, General Motors R&D, 30500, Mound Rd, Warren, MI 48092, USA. wende.zhang@gm.com

a bicyclist is tracked in subsequent video frames with an extended Kalman filter (EKF) based tracking algorithm which uses a simple point model and perspective projection for motion and measurement models, respectively.

The remainder of this paper is organized as follows. Section II reviews related work on detection and tracking of pedestrians. Our primary technical contributions in detection and tracking are described in Sections III and IV respectively. We describe experimental results using the system in Section V and conclude in Section VI.

II. RELATED WORK

There is a significant body of work on vision-based approaches (using detectors primarily sensitive to visible light) for pedestrian detection and tracking. Research using computer vision for pedestrian detection and tracking extends back a number of years. For a comprehensive survey of classical work, please see [9] and [11] while more recent work is surveyed in [8], [5], [4].

For the detection of pedestrians, there are roughly two main approaches: single template and part-based. This classification is based on representation of a human body regardless of features and classifiers used. Historically, a single template based approach was studied first and showed better performance compared to part-based models. Recently, however, some part-based models have shown more promising performance while they have a flexible and rich model. In a single template approach, the model captures a whole human body pattern using a single detection window. *Papageorgiou* et al. [15] uses Haar wavelet features in combination with a polynomial Support Vector Machine (SVM). *Viola* et al. [18] augment space-time information to their simple Haar-like wavelet features for moving people detection. *Dalal* and *Trigg* [3] show excellent performance for detecting human in a static image using a dense HOG (Histogram of Oriented Gradient) representation and a linear SVM. On the other hand, in a part-based approach, it captures the pattern of each part and then combines results to make a final decision for pedestrian detection. Generally, part-based approaches can handle with varying appearances of pedestrians due to clothing, pose, and occlusion, and thus, provide a more complex model for a pedestrian detection problem. *Mohan* et al. [15] divide human body into four parts: head, legs, left, and right arm. Each part detector is trained using a polynomial SVM and outputs are fed into a final classifier after checking geometric plausibility. *Mikolajczyk* et al. [14] model humans as assemblies of parts that are represented by the Scale Invariant Feature Transform (SIFT)-like orientation features. *Felzenszwalb* et al [7] demonstrate that a part-based model human detector can outperform many of existing current single template based approaches. Based on a variation of HOG features, they introduce a latent SVM formulation for training a part-based model from overall bounding box information without part location labelings.

For tracking of pedestrians, a number of mathematical frameworks have been proposed. Statistical or probabilistic methods such as the (extended) Kalman filter and particle filter are often employed. For instance, one such approach [10] uses an $\alpha - \beta$ filter to overcome gaps in detection where the proposed tracker is a simplified Kalman filter with a constant velocity model and predetermined steady-state gains. In another example [17], particle filters have been used to track a number of interacting people from a fixed camera. Other density estimation methods such as mean-shift [2] as well as structure from motion like optical-flow [16] have also been proposed.

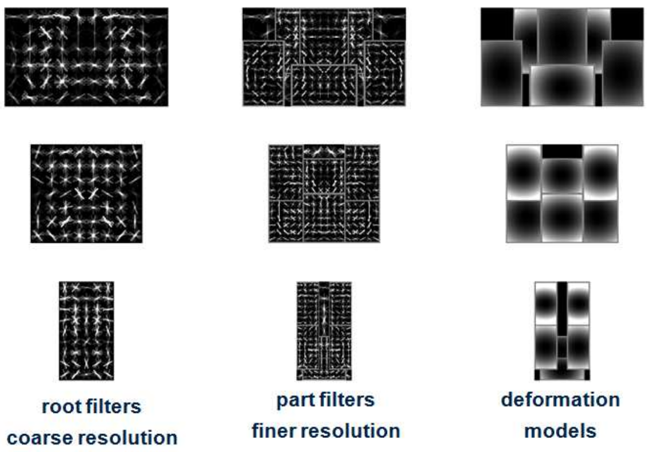


Fig. 2. Visualization of a three-component bicycle model. Each row corresponds to one specific view of a bicycle. Each column (from first to third) represents root filter, part filter, and deformation model, respectively.

III. BICYCLE DETECTION WITH A DEFORMABLE PART-BASED MODEL

Bicycle detection is challenging in that a bicycle presents dramatic appearance changes according to camera viewpoints and also has a intra-class variability (e.g., mountain bikes vs. racing cycles). One of the common solutions to tackle this problem is to establish part-based model for an object of interest. Rather than trying to capture a global pattern of an object with one template, part-based models focus on parts of an object and, in consequence, provide more flexible and robust representations. While part-based models have an elegant formulation in theory, they have not shown a better performance compared to a single template based approach. Recently, however, *Felzenszwalb* et al. [7] demonstrate a part-based model which outperformed the single template model by using a latent SVM formulation in combination with a variation of HOG features. In this paper, our work for bicycle detection is largely based on this work. The following subsections discuss some important details of the *Felzenszwalb* et al. [7] model and how it was applied to the algorithm in this research.

A. HOG Features

Selecting the correct feature is important because overall performance of the system depends on the discriminative power of features used in detection algorithm. As discussed in Section II, most successful features are the same regardless of the type of approach (i.e., single template or part-based). They are Haar wavelet, edgelet, and histogram based features such as SIFT [13] and HOG. Among those, the HOG feature has been considered as one of the strongest feature and used as a basic ingredient of more sophisticated feature sets. Basically, the HOG feature captures the shape information of an object and this aspect is naturally revealed via a visualization of HOG feature. Figure 2 illustrates HOG representation of some viewpoints of a bicyclist. According to the recent comprehensive evaluation studies [4] and [5], the HOG feature still shows best performance as a single feature relative to other existing feature sets. The deformable part-based model of *Felzenszwalb* et al. [7], which is our baseline detector, also uses HOG features as a building block. In fact, they use a PCA (Principal Component Analysis) version of HOG features. They report that a new 13-dimensional feature set obtained by performing PCA over an original 36-dimensional feature set can capture essentially the same information. This dimensionality reduction of features not only takes advantage of its highly discriminative power, but also speeds up the detection and training processes.

B. *Deformable Part-Based Model*

The core ideas of the deformable part-based model of *Felzenszwalb* et al. [7] can be summarized with three factors: a deformable part representation, an efficient matching process, and a latent SVM training process.

First, they define a star-structured part-based model which is composed of a root filter, n (usually six) part filters, and associated deformation parameters. A root filter is for capturing an overall shape of an object (shown in the first column in Figure 2) and part filters are for capturing the appearance of each part of an object (shown in the second column). Finally, deformation parameters are for measuring the deviation of the part from its ideal location (shown in the third column). Thus, the score of the star model at a particular position and scale is defined by the sum of root filter score and part filter scores (from the best possible placement of the parts) subtracted by a deformation cost. The authors also introduce a mixture of this star model to handle with significant changes in appearance according to viewpoint variation. Second, an efficient matching process based on dynamic programming and generalized distance transforms [7] is proposed. With the mixture of star models, since a matching process itself is a huge optimization problem, it is most important to incorporate a fast method for a detection task. Finally, a latent SVM training process is formulated to train a mixture of star models from bounding box ground truth. As the ground truth does not include part labeling information, part locations are treated as latent variables during training and thus the whole problem boils down to an optimization task with two sets of variables. In practice, they solve this problem using a coordinate descent algorithm by alternating between finding better latent values and optimizing the latent SVM objective function. In a detection process each example x is scored by a function of the following form:

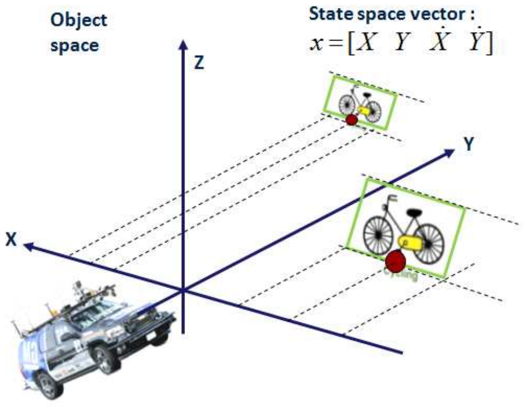


Fig. 3. Bicycle tracking problem formulation

$$f_\beta(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z). \tag{1}$$

where β is a vector of model parameters, z are latent values, and $\Phi(x,z)$ is a feature vector. In one star model, β is the concatenation of the root filter, the part filters, and deformation cost weights, z is a specification of the object configuration, and $\Phi(x,z)$ is concatenation of subwindows from a feature pyramid and part deformation features. We refer the reader to [7] for more details.

IV. BICYCLE TRACKING WITH AN EKF ALGORITHM

Once a bicycle detector is fired for the predefined number of times, the next step is to track its location from frame to frame. Because of the relatively high cost of the detector, we are interested in incorporating an algorithm with a lower complexity for tracking. For this reason, we chose to apply a traditional EKF (extended Kalman filter) to our framework. We assume that a bicycle has a simple point motion model with a constant velocity. In addition, as a measurement model, a nonlinear perspective projection equation is linearized and fed into the EKF framework. Specifically, the tracking is conducted via the following three steps:

- **Step 1:** Back-project a low midpoint in a bounding box (from detector) from the image coordinates into the vehicle coordinates.
- **Step 2:** Run the EKF prediction step to predict its next position using a simple point motion model.
- **Step 3:** Run the EKF update step to incorporate the detection results in the next frame and forward-project the point into the image coordinates again and update its bounding box.

We discuss technical details of both a motion model and a measurement model in the next subsections.

A. *Bicycle Point Motion Model*

Let's consider a bicycle tracking problem illustrated in Figure 3. Since a bicycle has its own unique kinematics, at a first glance, it seems natural to use a bicycle's kinematics as a motion model. However, it is a completely different situation once considering the measurement characteristics. The measurement in our case is a rough bounding box in the image space. From the sequence of these measurement, estimating all state variables (e.g., yaw and yaw rate) of the complicated model is a challenging task. We believe more comprehensive experiments are needed for this. As a starting point, we assume that a bicycle can be seen as a moving mass and thus, we use a simple point motion model for tracking. We use the midpoint of the bottom line of a bounding box (displayed as a red dot in Figure 3) as a representative point. Based on a flat ground assumption, the point can move freely only in the X-Y plane in vehicle coordinates. Thus, the state of this moving point on time step k is expressed as a vector:

$$\mathbf{x}_k = [x_k \quad y_k \quad \dot{x_k} \quad \dot{y_k}]^T \tag{2}$$

and the continuous-time state equation for this constant velocity model [1] can be modeled as a linear, time-invariant system:

$$\frac{d\mathbf{x}(t)}{dt} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{w}(t) \tag{3}$$

where $\mathbf{w}(t)$ is a continuous time white noise process. A discrete model of this state-space equation is used for the Kalman filter.

B. Bicycle Measurement Model

In our work, since only a monocular camera is used as a sensor device, the measurements are bounding box positions in the image space, which are results of the detection process. In addition, the tracking process itself is executed in the state space (i.e., in the vehicle coordinate). Thus, a measurement model should be able to map the state variable $\mathbf{x}$ into its measurement space (i.e., in the image coordinate) and this is done by a perspective projection equation. The nonlinear mapping of the state space into the measurement space of the video camera is given by:

$$\mathbf{y}_k = h(\mathbf{x}_k, k) + \mathbf{v}_k \tag{4}$$

where $\mathbf{v}_k$ is the measurement noise on the time step k and the nonlinear mapping function h is obtained by the following transformation:

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} f/s_x & 0 & u_c \\ 0 & f/s_y & v_c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{R} & | & \mathbf{t} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$
$$= \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{5}$$

where s_x and s_y are scale factors in x and y respectively, and (u_c, v_c) is a camera optical center and f is the focal length of a camera. $\mathbf{R}$ is a rotation matrix and $\mathbf{t}$ is a translation vector for extrinsic parameters. The parameters a_{ij} are the corresponding entries of the final perspective projection matrix. Based on a flat ground assumption, the vector function h is expressed by:

$$h_1 = \frac{a_{11}X + a_{13}Z + a_{14}}{a_{31}X + a_{33}Z + a_{34}} \qquad h_2 = \frac{a_{21}X + a_{23}Z + a_{24}}{a_{31}X + a_{33}Z + a_{34}} \tag{6}$$

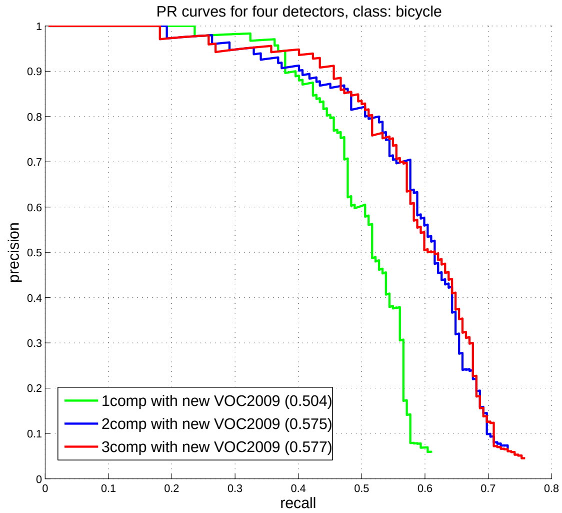


Fig. 4. PR curves for two detectors. The red plot shows the response for the three-component model trained with PASCAL2009 and our dataset and the blue plot and green plot show the response for two-component model and one-component model, recpectively, trained with the same dataset.

V. Experimental Results

We evaluated our detection and tracking framework using various real world datasets. We first conducted bicyclist detection experiments using the PASCAL VOC datasets [6] and a private dataset we collected from our experimental vehicle. With regard to bicycle tracking, we also collected video data of various scenarios in terms of bicycle movement. As the first set of data, six video sequences were recorded from a stationary vehicle. Tracking experiments for each case using the EKF algorithm are also conducted.

A. Performance Analysis of Detector

In bicycle detection experiments, we used 357 positive training samples and 3300 negative samples. Based on samples from the PASCAL VOC 2009 dataset `train` [1], we augmented the dataset with 160 positive samples from our private bicyclist dataset. We trained a three-component bicycle model which can capture three different viewpoints of a bicycle (i.e., frontal, 45°, and side view). Figure 2 visualizes this mixture model. For the test set, we used the same PASCAL's dataset `val` plus 100 test samples from ours. We

[1] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009/#devkit/, accessed on Sep. 5 2009

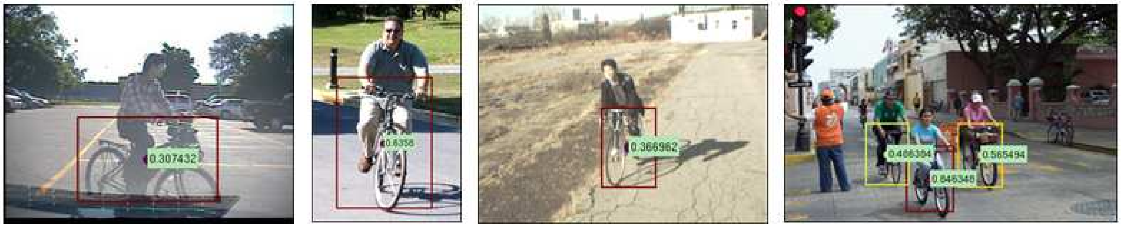
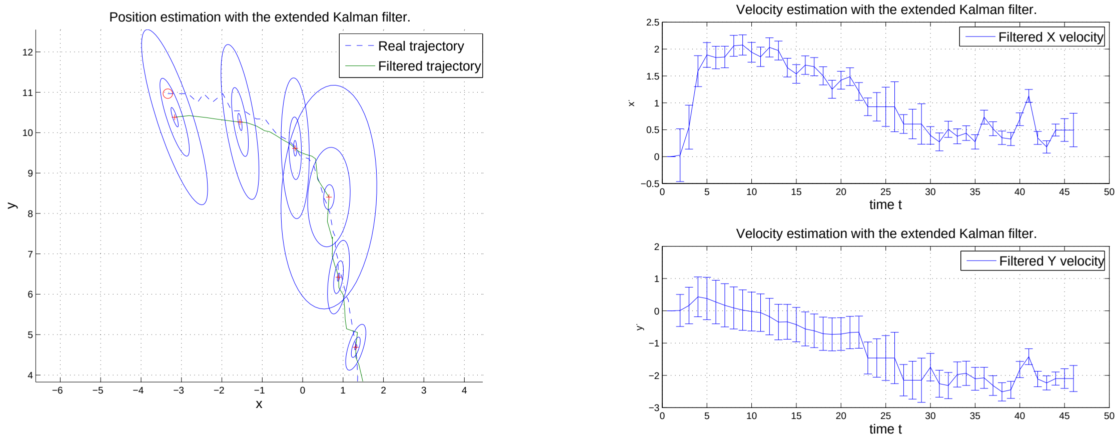


Fig. 5. Examples of detection with our three-component bicycle model



(a) The estimated path trajectory. The ellipses represent the 1, 2, 3-sigma confidence regions for the bicycle position.

(b) Illustration of tracking accuracy in sequence “stationary case”: the estimates of velocities of the bicycle in X and Y coordinates are plotted against frame numbers.

Fig. 6. Performance analysis of the tracker : stationary case

run the detector (equipped with our three-component bicycle model) over all images in the test set and draw a precision-recall (PR) curve for evaluation. A PR-curve for our model is compared with that of the two-component model of [7] as well as that of the one-component model in Figure 4 and some typical examples of detection are shown in Figure 5. We used the same evaluation criterion of VOC PASCAL competition for a detection task.

B. Performance Analysis of Tracker

Tracking experiments were conducted on the six videos collected from our experimental vehicle. All videos of a person riding a bicycle were recorded from the vehicles' cameras while the vehicle was stationary. We evaluate our EKF based tracking algorithm with a deformable part-based detector over all video sequences[2]. Here, due to the space limitation, we only analyze one specific scenario (‘sequence4’) in detail.

In the ‘sequence 4’ case, a bicyclist comes across the road in front of the vehicle and makes a turn toward the vehicle so that the left side and frontal view of the bicycle are seen and must be tracked. Table I illustrates the configuration of the image sequence.

TABLE I

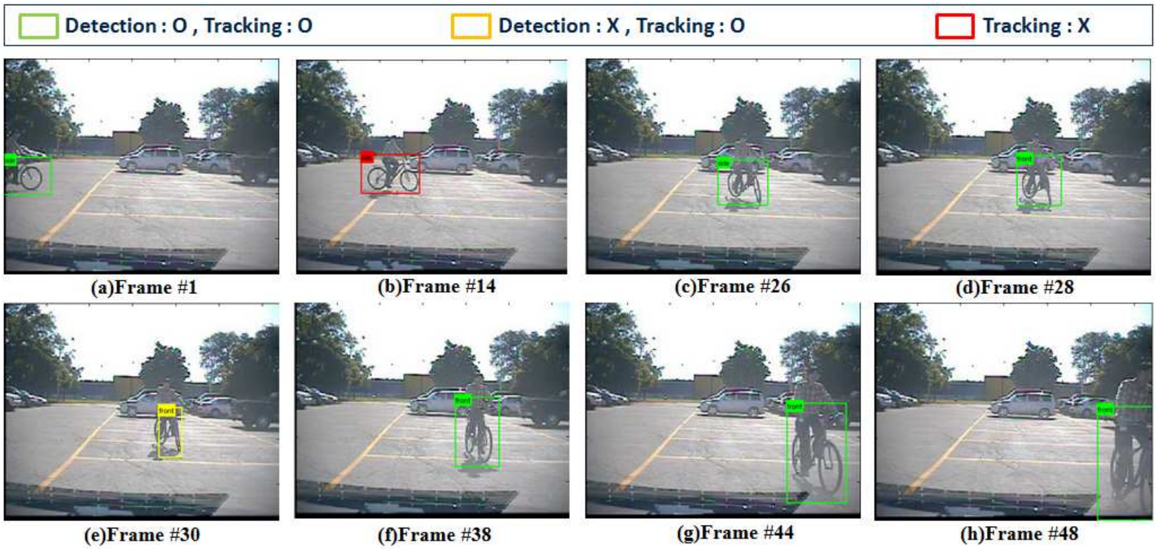
DETAILS OF ‘SEQUENCE 4’ USED IN THE EVALUATION

Sequences	Size	Frame-number	FPS	Bicyclist-number
‘sequence 4’	320×240	107	13	1

For the performance of tracking, as partially shown in Figure 7, the EKF based tracking algorithm successfully tracks the bicyclist except for the case in which the bicyclist is shown beyond the effective(or working) distance of our detector, which is around $10m$. More detailed analyses for each case are investigated by plotting filtered state variables of the tracker at each time step. In our case, these are positions and velocities in X and Y coordinates of the bicyclist. For instance, Figure 6(a) shows a bird-eye view of the bicyclist's position and Figure 6(b) shows the velocity estimation in both X and Y coordinates for the ‘sequence 4’ case (see Figure 6 for the details).

In terms of the speed of the detection and tracking, detection (approximately 0.3 s/frame on a P-IV 2GHz computer with 2GB memory) is much more time consuming than tracking (0.1 s/frame). The reason we put more computational complexity on the detector is that we believe good tracking

[2] The videos of tracking results over six sequences are available at http://www.cs.cmu.edu/h̃yunggic/tracking.



(a)Frame #1 (b)Frame #14 (c)Frame #26 (d)Frame #28

(e)Frame #30 (f)Frame #38 (g)Frame #44 (h)Frame #48

Fig. 7. Tracking results for bicyclist: stationary case ('sequence 4'). The green bounding box means reliable detection and reliable tracking. The yellow bounding box means unreliable detection and reliable tracking. Red bounding box means unreliable tracking.

performance can be achieved from good observations. Also, this approach is a main stream of current research trend which is called 'Tracking-by-Detection' or 'Detection-based Tracking'.

VI. Conclusions and Future Work

This paper presents a bicycle detection and tracking framework. To robustly detect bicycles, we applied *Felzenszwalb*'s deformable part-based detector [7] into our framework. Using the method, we build up a more powerful three-component bicycle model. We achieved a dramatic speed improvement for the detection process by exploiting known geometric constraints. Once the bicycle has been detected in the image, the object is tracked in subsequent video frames with an EKF based tracking algorithm which use simple point model and perspective projection for a motion model and a measurement model, respectively. This complementary approach allows our system to effectively track a bicyclist even when he/she changes orientations in the image. Several experiments shows the effectiveness of each component of the proposed framework. As part of our future work, we intend to develop an Interacting Multiple Model (IMM) based tracking algorithm which takes into account different bicycle motion kinematics at different traffic contexts.

VII. Acknowledgments

This project was funded by General Motors through the General Motors-Carnegie Mellon Autonomous Driving Collaborative Research Lab.

References

[1] Y. Bar-Shalom. Tracking and data association. 1987.

[2] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Conference on Computer Vision and Pattern Recognition*, 2005.

[4] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. *Conference on Computer Vision and Pattern Recognition*, 2009.

[5] M. Enzweiler and D. Gavrila. Monocular pedestrian detection: Survey and experiments. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2008.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results. http://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints), 2009.

[8] T. Gandhi and M. M. Trivedi. Pedestrian protection systems: Issues, survey, and challenges. *IEEE Transaction on Intelligent Transportation System*, 8(3):413–430, 2007.

[9] D. M. Gavrila. Sensor-based pedestrian protection. *IEEE Intelligent System*, 16(6):77–81, 2001.

[10] D. M. Gavrila, J. Giebel, and S. Munder. Vision-based pedestrian detection: The protector system. *IEEE Intelligent Vehicle Symposium*, pages 13–18, 2004.

[11] Z. Li, L. L. K. Wang, and F. Wang. A review on vision-based pedestrian detection for intelligent vehicles. *Conference on Vehicular Electronics and Safety*, 2006.

[12] J. Llinas and D. Hall. An introduction to multi-sensor data fusion. *International Symposium on Circuits and Systems*, 1998.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[14] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. *European Conference on Computer Vision*, I:69–81, 2004.

[15] C. Papageorgiou and T. Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33, 2000.

[16] J. Shi and C. Tomasi. Good features to track. *Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[17] K. Smith, D. Gatica-Perez, and J. M. Odobez. Using particles to track varying numbers of interacting people. *Conference on Computer Vision and Pattern Recognition*, 2005.

[18] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005.