Proceedings of the 2007 IEEE
Intelligent Transportation Systems Conference
Seattle, WA, USA, Sept. 30 - Oct. 3, 2007

Head Pose Estimation for Driver Assistance Systems: A Robust Algorithm and Experimental Evaluation

Erik Murphy-Chutorian, Anup Doshi, and Mohan Manubhai Trivedi
Computer Vision and Robotics Research Laboratory
University of California, San Diego
erikmc@ucsd.edu, andoshi@ucsd.edu, mtrivedi@ucsd.edu

Abstract— **Recognizing driver awareness is an important prerequisite for the design of advanced automotive safety systems. Since visual attention is constrained to a driver's field of view, knowing where a driver is looking provides useful cues about his activity and awareness of the environment. This work presents an identity- and lighting-invariant system to estimate a driver's head pose. The system is fully autonomous and operates online in daytime and nighttime driving conditions, using a monocular video camera sensitive to visible and near-infrared light. We investigate the limitations of alternative systems when operated in a moving vehicle and compare our approach, which integrates Localized Gradient Orientation histograms with support vector machines for regression. We estimate the orientation of the driver's head in two degrees-of-freedom and evaluate the accuracy of our method in a vehicular testbed equipped with a cinematic motion capture system.**

I. Introduction

With the increasing adoption of radar and laser-based sensor systems for adaptive cruise control and parking assistance, cars have gained the ability to detect the speed and position of obstacles in front or behind the vehicle. Currently these vehicle sensor systems operate in a purely nondiscriminatory manner, either constantly in operation, (e.g. to match the speed of a neighboring vehicle or to notify the driver of obstacles while moving in reverse), or turned off. These safety systems are unused during the remainder of the vehicle operation because they have no ability to operate in a context-specific manner. Consider the situation where a driver is making a right turn, but looking out his left window and unaware of a pedestrian that suddenly stepped in front of his car. The forward-looking radar system could detect the pedestrian, but alerting the driver to every potential hazard would be both distracting and bothersome. Instead, if the automobile were able to recognize when a driver had not noticed a possible hazard, it could trigger an alert only when these dangerous situations arise. This fusion of interior and exterior observations comprises an important new paradigm in advanced vehicular safety [1], [2]. In this paper, we focus on an important component specific to our example: estimating the orientation of a driver's head.

A driver's field-of-view can be reasonably approximated from the pose of his head, which can unobtrusively monitored by a small video camera. We present a new online system to estimate a driver's head pose from a single video camera during daytime or nighttime driving. To evaluate and compare our system, we simultaneously capture ground truth data with a marker-based motion capture system installed in the cabin of our test vehicle.

The wide variation in personal appearance – caused by differences in hair, skin-color, facial features, and shape complicates the design of an accurate Head Pose Estimation System (HPES). These problems are further compounded in a moving vehicle, where ever-shifting lighting conditions cause heavy shadows and illumination changes. As a result, techniques that demonstrate high proficiency in a laboratory setting are prone to failure when operated in an automobile.

In designing our system, we strove for an approach that met the following criteria:

1) **Monocular**: The system should be able to estimate head pose from a single camera. Although accuracy might be improved with stereo imagery, this would require additional hardware and periodic calibration, both of which are undesirable in a mass-production system.
2) **Autonomous**: There should be no required initialization, nor any concept of estimation drift over time. This precludes the use of pure-tracking approaches that measure the relative head pose with respect to some initial configuration.
3) **Invariant**: The system must work regardless of the specific driver and operating conditions.

A. Prior Work

There have been many innovative approaches to head pose estimation, and we limit this discussion to systems that satisfy our first two design considerations.

The most straightforward way to estimate head pose from a static image is to compare a new head view to a set of training examples (each labeled with a discrete pose) and find the most similar view. Systems have been proposed that compare these views using normalized cross-correlation [3], mean squared error [4], differences in gradient direction [5], elastic graph matching [6], and distance between subspace projections [7], [8]. These *prototype methods* are appealing because training requires only positive examples, but their winner-take-all pose estimation can be greatly affected by a single noisy measurement.

Other approaches use machine learning techniques to train an array of face detectors each attuned to a specific pose

1-4244-1396-6/07/$25.00 ©2007 IEEE.

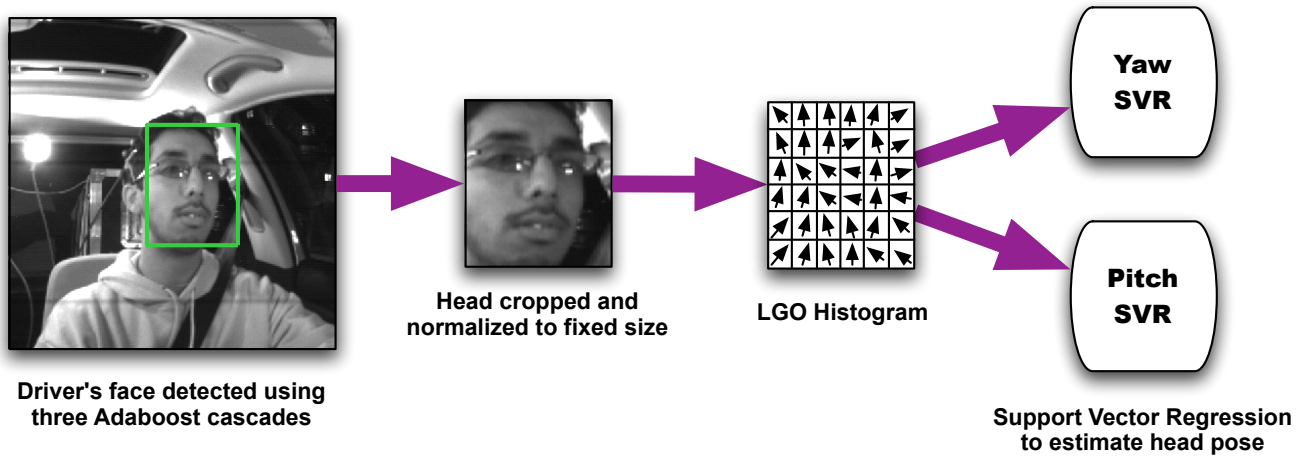

Fig. 1. An overview of our proposed Head Pose Estimation System

direction. These *detector array* systems have consisted of neural networks [9], Adaboost cascades [10], view-based subspace energy detectors [11], and support vector classifiers in a kernel principal component analysis subspace [12]. These systems provide excellent invariance to identity, but due to the complexity of training many detectors, they have typically been limited to fewer than 10 discrete poses, varying across a single degree of freedom.

Non-linear regression techniques can also be used to directly project an image into the space of pose variation. This has been accomplished with locally-linear maps [13], support vector regressors [14], locally-linear projection [15], and locally embedded analysis [16]. In these *manifold projection* methods, pose is estimated as a continuous measurement – typically constrained to two degrees-of-freedom (pitch and yaw).

In addition, there are a few systems to note for their application of head pose estimation in an automobile [5], [17]–[20].

In this paper, we present a new HPES that is capable of online operation in a car. Our approach uses soft histograms of location-specific gradient orientation as the input to a support vector regressor for each degree-of-freedom. We compare our method to a system that uses support vector regression applied to a feature vector created by principal component analysis of raw image gradients [14]. Our method exhibits a significant improvement in estimation accuracy, and maintains performance in the variable automotive setting. In addition, we also compare our system to a normalized cross-correlation-based prototype method to advocate the use of our more complex approach.

In Section II we present our system for head pose estimation, and in Section III we describe our experimental testbed. Section IV contains our evaluation and comparison to existing techniques, and in Section V we provide our concluding remarks.

II. System Description

This section describes each of the following HPES subcomponents in detail.

1) Facial regions are found by three cascaded-Adaboost face detectors applied to the grayscale video images;
2) The detected facial region is scale-normalized to a fixed size and used to compute a Localized Gradient Orientation (LGO) histogram;
3) The histogram is passed to two Support Vector Regressors (SVRs) trained for head pitch and yaw.

A graphical overview can be found in Figure 1.

A. Facial Region Detector

To detect the location of the driver's head, we use three Adaboost cascades attuned to left profile, frontal, and right profile faces [21], [22]. Each detector is capable of recognizing heads with enough deviation from its characteristic pose that when combined, they reasonably span the range of head poses in our training data: $-30°$ to $20°$ in pitch and $-80°$ to $80°$ in yaw. For both training and testing, an uncompressed, grayscale image is used as the input to the detectors, and we consider the largest detected rectangular region as the location of the driver's face. In our automobile experiments, this facial detection scheme successfully detected a region in 87% of the video frames.

To ensure that the system is invariant to scale, every region is down-sampled to a fixed size of 34x34 pixels. In an automobile, this makes the system invariant to the distance between the driver and the camera, which varies between drivers.

B. Localized Gradient Orientation Histogram

To provide a robust description of each facial region, we compute a soft histogram of localized gradient orientation. This representation was first presented as part of the Scale-Invariant Feature Transform [24], intended for correspondence matching between regions surrounding scale- and rotation-invariant keypoints. This histogram provides a compact feature representation that is robust to minor deviations

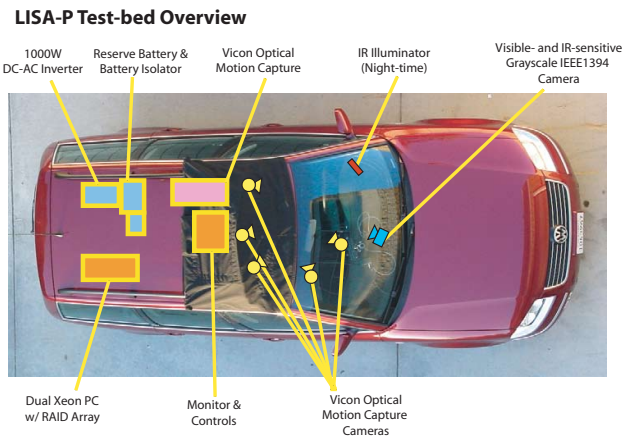

Fig. 2. The LISA-P Experimental Testbed used for data collection and evaluation [23].

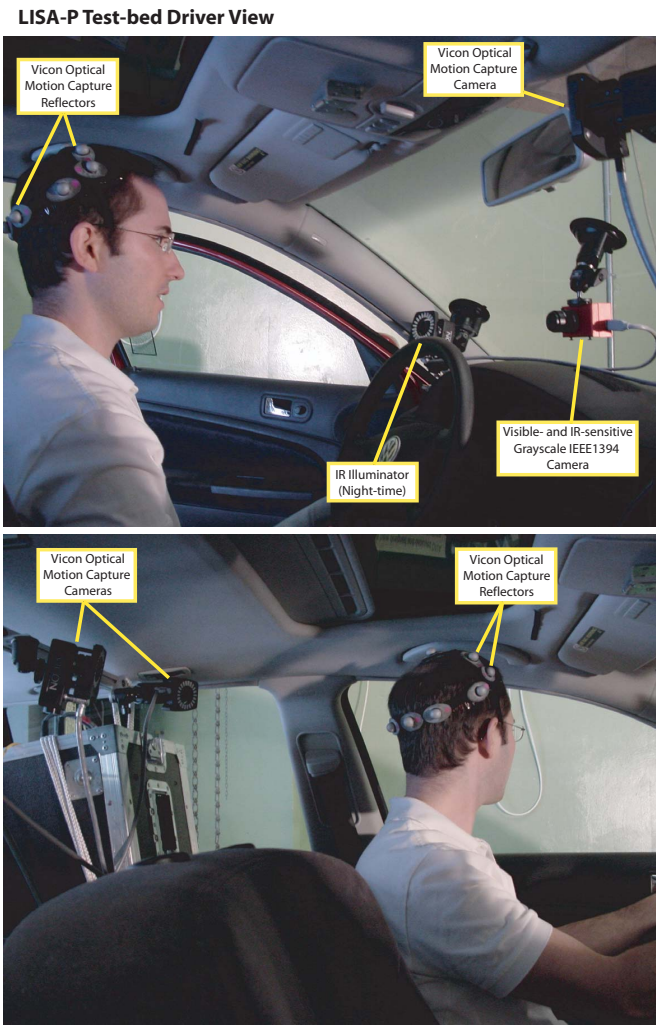

Fig. 3. LISA-P Experimental Testbed: Driver view. These images depict the experimental setup used for to simultaneously capture video of the driver while collecting accurate head pose information.

in region alignment, lighting, and shape [24], [25]. Although commonly referred to as a SIFT descriptor, this histogram-based descriptor is entirely separate from the SIFT procedure for finding scale- and orientation-invariant keypoints. Instead, we call this representation a Localized Gradient Orientation (LGO) histogram, applying the term to describe the SIFT descriptor when generalized for any spatial window size and any number of histogram bins.

In contrast to object recognition systems that represent an object as a configuration of multiple histogram descriptors [24], we use a single LGO histogram to represent the entire scale-normalized facial region. This descriptor consists of a three-dimensional histogram where the first two dimensions corresponding to vertical and horizontal positions in the image, and the third to the gradient orientation. For an $M \times N \times O$ histogram, let the triplet (m, n, o) denote a specific bin in the histogram. The horizontal and vertical image gradients, $\boldsymbol{X}_x(x, y)$ and $\boldsymbol{X}_y(x, y)$, are approximated by filtering with 3×3 pixel Sobel kernels. The image is then split into $M \times N$ discrete blocks, and for each pixel (x, y) in the (m, n) block, the absolute gradient orientation $o_{x,y}$ is quantized into one of O discrete levels,

$$o_{x,y} = \left\lfloor O \times \left(\frac{1}{2\pi} \operatorname{atan2}\left(\boldsymbol{X}_y(x, y), \boldsymbol{X}_x(x, y)\right) + 0.5 \right) \right\rfloor, \quad (1)$$

and used to increment the $(m, n, o_{x,y})$ histogram bin. After computing the histogram, it is smoothed with the $3 \times 3 \times 3$ kernel,

$$K(m, n, o) = \left(1 - \frac{g(m)}{M}\right)\left(1 - \frac{g(n)}{N}\right)\left(1 - \frac{g(o)}{O}\right), \quad (2)$$

to prevent aliasing effects, for $\{m, n, o \in (-1, 0, -1)\}$, where $g(\cdot)$ is the complement impulse function,

$$g(n) = \begin{cases} 0 \text{ if } n = 0 \\ 1 \text{ if } n \neq 0 \end{cases}. \quad (3)$$

The resulting *soft* histogram is subsequently reshaped and normalized to a unit vector. Finally, as suggested by Lowe [24], any vector component greater than 0.2 is truncated to 0.2, and the vector is renormalized if necessary. In our system, we use a 128-dimensional vector, where $M = 4$, $N = 4$, and $O = 8$.

C. Support Vector Regression

Support vector regression is a supervised learning technique for nonlinear regression of a scalar function [26], [27]. The basic concept is to use a nonlinear kernel function to project the input data into a high-dimensional space and then use linear regression to fit a hyperplane

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \Phi(\boldsymbol{x}) - b, \quad (4)$$

in this space. This is accomplished by simultaneously flattening the hyperplane, e.g. minimizing $\|\boldsymbol{w}\|^2$, while also minimizing the sum of the error from data points which lie outside a margin, ϵ, surrounding the hyperplane. Explicitly projecting the input data into a high- or infinite-dimensional space is computationally infeasible, however, a fundamental aspect of a support vector machine is that every high-dimensional projection is part of a dot product with another

TABLE I

A COMPARISON OF MEAN ABSOLUTE ERROR BETWEEN HEAD POSE ESTIMATION APPROACHES.

Method	Laboratory		Daytime Driving		Nighttime Driving	
	Pitch	*Yaw*	*Pitch*	*Yaw*	*Pitch*	*Yaw*
Normalized Cross-correlation	**5.19**°	21.25°	11.30°	14.90°	5.94°	16.49°
Gradient PCA + SVR	5.72°	13.46°	**3.62**°	12.19°	5.19°	13.11°
Proposed System: LGO Histogram + SVR	5.58°	**6.40**°	3.99°	**9.28**°	**5.18**	**7.74**°

vector [28]. This conveniently saves us from ever calculating any high-dimensional projection by using only nonlinear kernel functions of the form,

$$\mathcal{K}(\boldsymbol{a}, \boldsymbol{b}) = \Phi(\boldsymbol{a}) \cdot \Phi(\boldsymbol{b}). \tag{5}$$

We use an optimized software package for support vector regression [29], and in our system we train the regressor with a radial basis function kernel,

$$\mathcal{K}_{RBF}(\boldsymbol{a}, \boldsymbol{b}) = \exp\left(-\gamma \|\boldsymbol{a} - \boldsymbol{b}\|^2\right). \tag{6}$$

We use two regressors, one trained for head pitch, and one for yaw. The input to both is the LGO histogram described in Section II-B. To find the optimum regression parameters ϵ and γ, as well as c, a parameter which adjusts the relative cost of data points that lie outside the margin, we scale-normalize each component of the training input, such that it spans the range [-1,1] and perform a cross-validation grid search across the parameter space. During testing, we use the same scaling parameters to normalize the new input before predicting the new pose.

III. EXPERIMENTAL SETUP

A. LISA-P Experimental Testbed

The LISA-P Experimental Test-bed, as seen in Figure 2, is used to collect the real-world test data. The vehicle is a modified Volkswagen Passat; images of its interior can be found in Figure 3. It is instrumented with a Vicon optical motion capture system, with 5 sensors placed in various locations around the driver's head. This marker-based system is used to gather precise ground truth head pose data [30].

The IEEE1394 camera used to capture face data is sensitive to both visible and near-infrared wavelengths, and is mounted on the windshield just under the rear-view mirror.

Additionally, for the purpose of illuminating the driver's face and stabilizing the lighting conditions at nighttime, a near-infrared illuminator is placed on the left-most part of the windshield. Since the emitted light is of not part of the visible spectrum, it does not serve as a distraction or cause any glare for the driver.

B. Head Pose Training

To train our system, we captured data using the Vicon Motion capture system. For the laboratory experiments, ten people were asked to sit on a chair against a white background while facing an IEEE1394 video camera. Behind the camera, a projector displayed a grid of points each representing a specific pose at 5° intervals spanning $(-30°, 20°)$ in pitch and $(-80°, 80°)$ in yaw. A cursor, corresponding to the subjects' current head pose, would follow the head position of the subject as they turned their head. When they passed over each of the 363 grid point locations for the first time, the point would change color and the Vicon system would capture an image of the subject. In this fashion we obtained a uniform sampling of all 10 subjects across the pose space. For the automotive experiments, we trained the system in a cross-validation scheme by extracting a uniform sampling of the pose space from five out of six subjects. This training procedure was repeated for every all-but-one combination, leaving the remaining subject for evaluation.

IV. SYSTEM EVALUATION

We compare our HPES in the laboratory and automobile to two other approaches. The first is a prototype matching scheme that uses normalized cross-correlation to compare the driver's face to each of the views in our training data. To make the system more robust to noise, we take the mean of the cross-correlation score for all training images that share the same discrete pose, and we estimate the head pose as the pitch and yaw corresponding to the maximum score after bi-cubic interpolation.

The second comparative HPES is our implementation of the Gradient-PCA system described by Li et al. [14]. We chose this work for comparison since it is the most similar to our proposed system, and it is capable of high accuracy and speed. This approach also uses two support vector regressors to estimate pitch and yaw. Instead of LGO histograms, the input to each regressor is the raw horizontal and vertical image gradient reduced to a 50-dimensional vector using principal component analysis. The PCA basis is derived from the training data.

For both of these comparative approaches we use the same array of Adaboost cascades described in Section II-A to locate and normalize the the region of the image corresponding to the driver's face.

In the automotive evaluation, six subjects were asked to drive the LISA-P vehicle on a 12-minute drive, two during daylight hours and four at night. During this time, video frames were captured at 30 frames-per-second along with the true head position as recorded by the motion capture system, and we evaluated our system based on a 60-second interval from the middle of each of the six drives. To maintain an even distribution of head pose direction, we further reduced this set to provide a uniform sampling of pitch and yaw, resulting in 5558 video frames in our test set.

The results of these experiments can be found in Table I. Here we quantify each approach by the mean absolute error in pitch and yaw between the motion capture reference and the estimated orientation. In the laboratory experiment, the

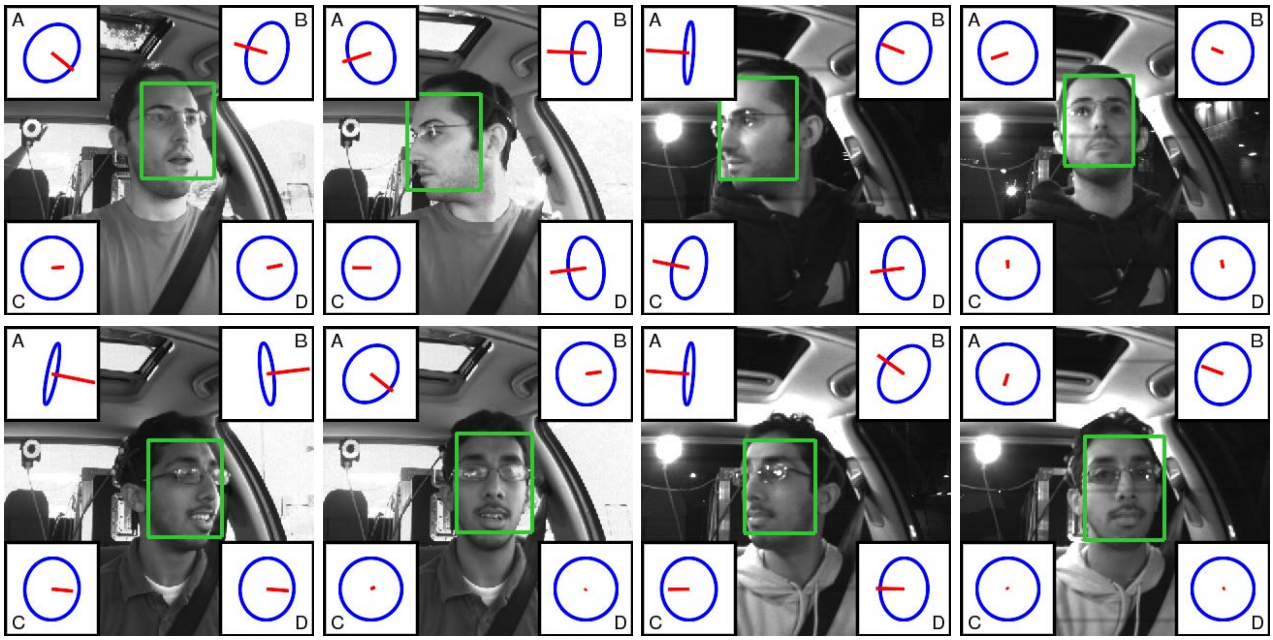

Fig. 4. A comparison of head pose estimation using the following methods: A – Normalized Cross-correlation prototype matching, B – Gradient PCA with support vector regression [14], C – LGO histograms with support vector regression, D – Vicon motion capture ground truth. The center square indicates the detected facial region using a trio of cascaded-Adaboost face detectors, and the pose for each method is indicated by the direction of the thumbtack.

LGO histograms provide a comparable level of error in pitch, and they outperform the Gradient-PCA approach with a $7.06°$ reduction in yaw error. In the driving experiments, our system outperforms the other approaches in absolute yaw error by a significant margin: $9.28°$ compared to $14.90°$ and $12.19°$ during the daytime experiment, and $7.74°$ compared to $16.49°$ and $13.11°$ during the nighttime drives. Examples of all three systems along with the ground truth data are presented in Figure 4.

V. Conclusions

Robust systems for driver activity monitoring will play a key role in the development of advanced driver assistance systems. In combination with environmental sensors, cars can be designed with the ability to supplement the driver's awareness to preempt and prevent hazardous situations. In this work we focused on an automotive HPES, since head pose is a strong indicator of a driver's field-of-view and current focus of attention. Head pose estimation is intrinsically linked to visual *gaze estimation*, the ability to characterize the direction in which an eye is focused. Alone, head pose provides a coarse indication of gaze, and one that can be estimated in situations where the eyes of a person are not visible (such as low-resolution imagery, or in the presence of eye-occluding objects like sunglasses). The addition of eye-gaze information would provide a better indication of gaze direction, but eye-gaze is only meaningful in conjunction with head pose information, as a persons prediction of gaze comes from a combination of both head pose and eye direction [31]. The presence of this effect suggests that humans are unable to estimate the true orientation of an eye, but instead able estimate the position of the eye relative to the position of the head. This effect is also present in many commercial eye trackers that require the subject to maintain a frontal head position.

The primary contribution of this paper is a new approach and evaluation that overcomes the difficulties that arise with varying lighting conditions in a moving car. Our approach uses Localized Gradient Orientation histograms to tolerate deviations caused by scale, position, rotation, and lighting. Using this representation, we provide a stable input to a support vector regressor for robust head pose estimation in two degrees-of-freedom. Our unoptimized system runs at approximately 5 frames-per-second, limited primarily by the time required to process the Adaboost cascades.

The HPES described in this paper is a stand-alone approach that operates on a single video frame. Greater accuracy and stability could be obtained by combining this framework with a visual tracking system. Pose would be estimated in each frame, validating the tracked configuration and re-initializing the tracker if necessary. This type of combination has been quite successful [32], and we believe this improvement would similarly extend to our automotive HPES.

Acknowledgments

We thank the Volkswagen Electronics Research Laboratory and the U.C. Discovery program for supporting for this research. In addition, we thank our team members in the Computer Vision and Robotics Research laboratory for their participation as test subjects and for their helpful suggestions and assistance. Mr. Shinko Cheng's contributions in the design and development of the LISA-P testbed and motion

capture system are gratefully acknowledged.

REFERENCES

[1] M. Trivedi, T. Gandhi, and J. McCall, "Looking-in and looking-out of a vehicle: Computer-vision-based enhanced vehicle safety," *IEEE Trans. Intelligent Transportation Systems*, vol. 8, no. 1, pp. 108–120, 2007.

[2] S. Cheng and M. Trivedi, "Holistic sensing and dynamic active displays," *IEEE Computer*, vol. 40, no. 5, pp. 60–68, 2007.

[3] D. Beymer, "Face recognition under varying pose," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1994, pp. 756–761.

[4] J. Ng and S. Gong, "Composite support vector machines for the detection of faces across views and pose estimation," *Image and Vision Computing*, vol. 20, no. 5-6, pp. 359–368, 2002.

[5] R. Pappu and P. Beardsley, "A qualitative approach to classifying gaze direction," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 1998, pp. 160–165.

[6] N. Krüger, M. Pötzsch, and C. von der Malsburg, "Determination of face position and pose with a learned representation based on labelled graphs," *Image and Vision Computing*, vol. 15, no. 8, pp. 665–673, 1997.

[7] J. Sherrah, S. Gong, and E.-J. Ong, "Face distributions in similarity space under varying head pose," *Image and Vision Computing*, vol. 19, no. 12, pp. 807–819, 2001.

[8] J.-W. Wu and M. Trivedi, "An integrated two-stage framework for robust head pose estimation," in *Proc. IEEE Int'l. Workshop Analysis and Modeling of Faces and Gestures*, 2005, pp. 321–335.

[9] H. Rowley, S. Baluja, and T. Kanade, "Rotation invariant neural network-based face detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1998, pp. 38–44.

[10] M. Jones and P. Viola, "Fast multi-view face detection," Mitsubishi Electric Research Laboratories, Tech. Rep. 096, 2003.

[11] S. Srinivasan and K. Boyer, "Head pose estimation using view based eigenspaces," in *Proc. Int'l. Conf. Pattern Recognition*, 2002, pp. 302–305.

[12] S. Li, Q. Fu, L. Gu, B. Scholkopf, Y. Cheng, and H. Zhang, "Kernel machine based learning for multi-view face detection and pose estimation," in *Proc. IEEE Int'l. Conf. Computer Vision*, 2001, pp. 674–679.

[13] R. Rae and H. Ritter, "Recognition of human head orientation based on artificial neural networks," *IEEE Trans. Neural Networks*, vol. 9, no. 2, pp. 257–265, 1998.

[14] Y. Li, S. Gong, and H. Liddell, "Support vector regression and classification based multi-view face detection and recognition," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 2000, pp. 300–305.

[15] B. Raytchev, I. Yoda, and K. Sakaue, "Head pose estimation by nonlinear manifold learning," in *Proc. Int'l. Conf. Pattern Recognition*, 2004, pp. 462–466.

[16] Y. Fu and T. Huang, "Graph embedded analysis for head pose estimation," in *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, 2006, pp. 3–8.

[17] Y. Zhu and K. Fujimura, "Head pose estimation for driver monitoring," in *IEEE Intelligent Vehicle Symposium*, 2004, pp. 501–506.

[18] J. Wu and M. Trivedi, "Visual modules for head gesture analysis in intelligent vehicle systems," in *IEEE Intelligent Vehicle Symposium*, 2006, pp. 13–18.

[19] E. Murphy-Chutorian and M. Trivedi, "N-tree disjoint-set forests for maximally stable extremal regions," in *Proc. British Machine Vision Conference*, 2006, pp. 739–748.

[20] K. Huang and M. Trivedi, "Driver head pose and view estimation with single omnidirectional video stream," *Automation and Remote Control*, vol. 25, pp. 821–837, 2006.

[21] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[22] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *IEEE Int'l. Conf. Image Processing*, vol. 1, 2002, pp. 900–903.

[23] S. Cheng and M. Trivedi, "Turn-intent analysis using body pose for intelligent driver assistance," *IEEE Pervasive Computing*, vol. 5, no. 4, pp. 28–37, 2006.

[24] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[25] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[26] H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems*, 1996, pp. 155–161.

[27] A. Smola and B. Scholkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[28] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. Annual Conf. Computational Learning Theory*, 1992, pp. 144–152.

[29] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[30] A. D. Wiles, D. G. Thompson, and D. D. Frantz, "Accuracy assessment and interpretation for optical tracking systems," in *SPIE Proceedings on Medical Imaging, Visualization, Image-Guided Procedures, and Display*, no. 5367, 2004, pp. 421–432.

[31] S. Langton, H. Honeyman, and E. Tessler, "The influence of head contour and nose angle on the perception of eye-gaze direction," *Perception and Psychophysics*, vol. 66, no. 5, pp. 752–771, 2004.

[32] T. Jebara and A. Pentland, "Parametrized structure from motion for 3d adaptive feedback of faces," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 144–150.