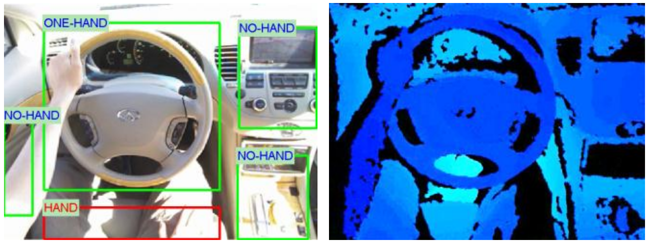
The Power is in Your Hands:
3D Analysis of Hand Gestures in Naturalistic Video



Eshed Ohn-Bar and Mohan M. Trivedi
Computer Vision and Robotics Research Lab
University of California at San Diego

June 26, 2013

Acknowledgments

UC Discovery Program and Associated Industry Partners

KETI (Korea Electronics Technology Institute

NextChip Inc.

Toyota Motor Corporation CSRC Initiative

Members of the LISA and CVRR lab for their help

Presentation Outline

- Motivation and Background
- Case Study
  Looking at Driver's Hands, Issues and Challenges
- Hand Detection - Integrated Cues Analysis
- Experimental Studies and Evaluations
- Concluding Remarks and Directions

(taken from NHTSA's analysis of driver inattention using a case-crossover approach on 100-car data: final report)

Vision-based Hand Activity Recognition

Simple
(lab settings)

A. Kurakin *et al.*, EUSIPCO 2012

M. Van den Bergh *et al.*, WACV 2011

Vision-based Hand Activity Recognition

Simple
(lab settings)

A. Kurakin *et al*., EUSIPCO 2012

M. Van den Bergh *et al*., WACV 2011

Hand detection was mostly studied in **indoor settings**, where the hand is segmented in a naive manner.

1) the hands may be the main salient object in the scene in terms of motion
2) skin-color
3) or it may be segmented using a depth-based threshold.

As single cues, such techniques were shown to **perform poorly** on our dataset

Vision-based Hand Activity Recognition

Simple
(lab settings)

A. Kurakin *et al*., EUSIPCO 2012

M. Van den Bergh *et al*., WACV 2011

Complex

A. Mittal, A. Zisserman, and P.H.S Torr, BMVC 2011

Hand detection was mostly studied in **indoor settings**, where the hand is segmented in a naive manner.

1) the hands may be the main salient object in the scene in terms of motion
2) skin-color
3) or it may be segmented using a depth-based threshold.

As single cues, such techniques were shown to **perform poorly** on our dataset

Motivation

1) Study recognition of naturalistic driver hand and hand +object gestures that are related to driver attention and intentions.

2) Thorough study of different feature extraction methods

3) Emphasis on robustness: Integrating models and cues from each region using a second-stage classifier.

(taken from NHTSA's analysis of driver inattention using a case-crossover approach on 100-car data: final report)

Case Study – In-Vehicle Hand Activity Recognition

Question: How often, if ever, do you perform each of the following actions while you are driving?	Ever	Often / Sometimes
Eat food and/or drink beverages	86%	57%
Have a long or serious discussion with a passenger	81%	49%
Talk on a cell phone while using the handset, not a hands free device	59%	27%
Talk on a cell phone while using a hands -free device	43%	27%
Set or change a GPS or direction finder	41%	21%
Send or receive text messages	37%	18%
Read a map	36%	10%

T. H. Poll, "Most U.S. drivers engage in 'distracting' behaviors: Poll," Insurance Institute for Highway Safety, Arlington, Va., Tech. Rep. FMCSA-RRR-09-042, Nov. 2011.

Motivation

Motivation

Simple Secondary Tasks	Moderate Secondary Tasks	Complex Secondary Tasks
Adjusting radio	Talking/Listening to Hand-Held Device	Dialing a hand-held device
Drinking	Inserting/Retrieving CD	Locating/Reaching/Answering Hand-Held Device
	Reaching for object (not hand-held device)	Operating/Viewing a PDA
	Eating	Reading

Percent of Secondary Task use in Crash/Near-Crash

Simple	Moderate	Complex	None	Total
24.4%	22.8%	4.6%	48.1%	100%

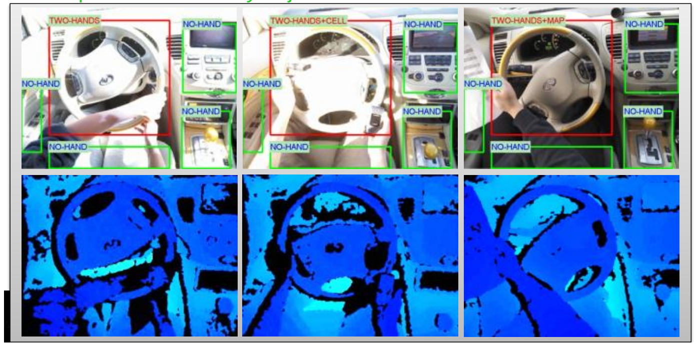
(taken from NHTSA's analysis of driver inattention using a case-crossover approach on 100-car data: final report)

CVRR-Hands 3D Dataset

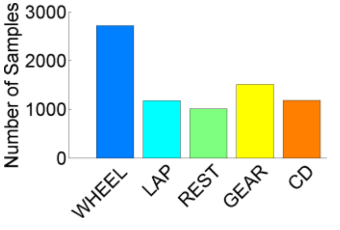
Complex – A unique effort compared to existing datasets and evaluations

- Hand-object Interactions
- Background Clutter
- Illumination Changes
- Frequent Occlusion by objects or other hand and self-occlusion.

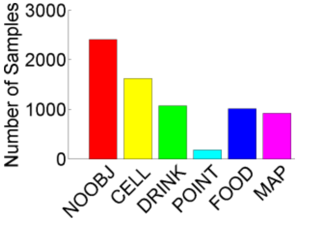
- **Over an hour of video in total**
- **Cross-subject testing**
- ***cvrr.ucsd.edu/eshed***



CVRR-Hands 3D Dataset



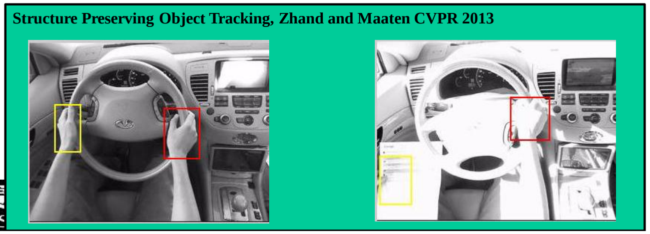
Hand Localization



Hand+Object Activities

LISA: Laboratory for Intelligent & Safe Automobiles

CVRR-Hands 3D Dataset



Proposed Approach – Integration of Regions

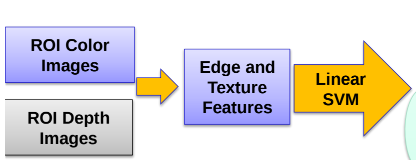
Observation: hand presence in a certain region can be detected, but the **difficult visual settings** make sliding-window detectors over the entire image perform poorly with **many false positives**.

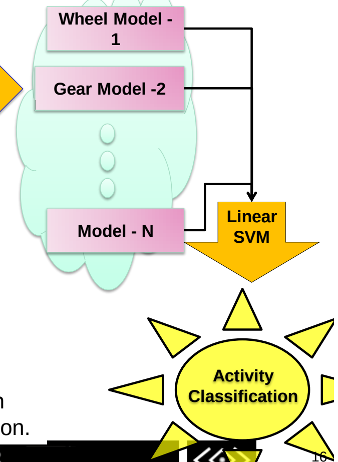
Proposed Approach – Integration of Regions

Observation: hand presence in a certain region can be detected, but the **difficult visual settings** make sliding-window detectors over the entire image perform poorly with **many false positives**.

Proposed Solution: **Constrain** to a number of regions useful for studying driver's state. Integrate cues from each region to perform final activity recognition.

Proposed Approach – Integration of Regions





Observation: hand presence in a certain region can be detected, but the **difficult visual settings** make sliding-window detectors over the entire image perform poorly with **many false positives**.

Proposed Solution: **Constrain** to a number of regions useful for studying driver's state. Integrate cues from each region to perform final activity recognition.

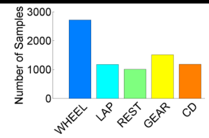
Proposed Approach – Integration of Regions – Why?

1) Hand and objects vary in appearance among the regions → requiring a *unique set of descriptors and a separate model*.

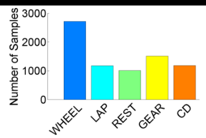
2) Hands are found in a subset of regions → *reducing the complexity of the detection problem in the entire scene.*

3) Each region, with different size and location, produce different challenges for a vision-based system → Leveraging several regions results in a higher-level reasoning of the hand configuration. Some regions may be more prone to illumination or larger in size, while others may require finer-detailed descriptors as a part of the arm may be present in them while the hand is interacting in a different region.

We Study Unbalanced Datasets



We Study Unbalanced Datasets



- Want to preserve all the variation in training

- One possible way to address this is through penalizing parameters in the SVM formulation so that the optimization problem is modified to be (biased penalties SVM)

$$\min_{w,b,\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C^+ \sum_{t_i=1} \xi_i + C^- \sum_{t_i=-1} \xi_i$$

$$\text{subject to} \quad t_i(\mathbf{w}^T\phi(x_i)+b) \geq 1-\xi_i, \quad \xi_i \geq 0,\ i=1,\ldots,l.$$

Vapnik, Statistical Learning Theory 1998

Features

Descriptor	Extraction Time (ms)	Descriptor Size
HOG99	6	11780D
MHOG11	10	9D
HOF88	13	1155D
DIFFHOG	10	9D
GIST8	370	2048D
Skin	10	4D
EUC	4	14535D
GLOBAL	1	3D

- HOG [1] and a ***modified HOG descriptor***.

- GIST [3].

- Skin:

A **color likelihood classifier** is constructed in the L*a*b color space for each user using an initialization frame. The final descriptor is composed of the area and area/perimeter ratio of the two largest connected components in the image.

- HOF: Haar-like on the optical-flow image and histogram

- GLOBAL: The median, mean, and variance of the intensities in the image. (Differ based on object or hand presence).

[1] N. Dalal and B. Triggs, IEEE Conference Computer Vision and Pattern Recognition, 2005
[2] N. Dalal ,B. Triggs, and C. Scmid, European Conference on Computer Vision, 2006
[3] A. Oliva and A. Torralba, International Journal of Computer vision, 2001

Features

- HOG [1] and a *modified HOG descriptor*.

- GIST [3].

- Skin:

A **color likelihood classifier** is constructed in the L*a*b color space for each user using an initialization frame. The final descriptor is composed of the area and area/perimeter ratio of the two largest connected components in the image.

- HOF: Haar-like on the optical-flow image and histogram

- GLOBAL: The median, mean, and variance of the intensities in the image. (Differ based on object or hand presence in depth image).

Descriptor	Extraction Time (ms)	Descriptor Size
HOG99	6	11780D
MHOG11	10	9D
HOF88	13	1155D
DIFFHOG	10	9D
GIST8	370	2048D
Skin	10	4D
EUC	4	14535D
GLOBAL	1	3D

IN MATLAB

[1] N. Dalal and B. Triggs, IEEE Conference Computer Vision and Pattern Recognition, 2005
[2] N. Dalal ,B. Triggs, and C. Scmid, European Conference on Computer Vision, 2006
[3] A. Oliva and A. Torralba, International Journal of Computer vision, 2001

A Modified HOG Feature Extraction

=> Gradient image (magnitude G and orientation θ)
=> Break into cells with 50% overlap
=> Orientation histogram for each cell

$$h^S(q) = \sum_{x,y \in S} G^S_{x,y} \cdot 1[\Theta(x,y) = \theta]$$

=> Concatenate

The *parameters* are the *number of cells* in the x and y direction in the entire region, and number of *orientation bins*.

Example:
2 × 2 grid of cell with 8 histogram bins results in a 32D feature vector.

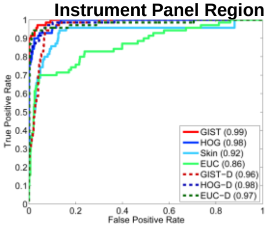
Final spatial descriptor for each time step:

$$\mathrm{h_t} = mHOG(I) = [h^1 \ldots h^{M \cdot N}]$$

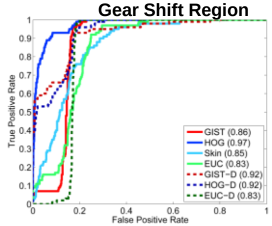
UCSD LISA: LABORATORY FOR INTELLIGENT & SAFE AUTOMOBILES CVRR

Results – Hand Detection in Single Regions

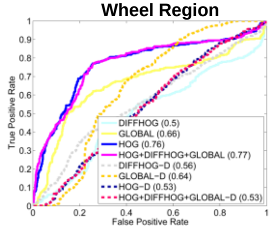
Instrument Panel Region



Gear Shift Region

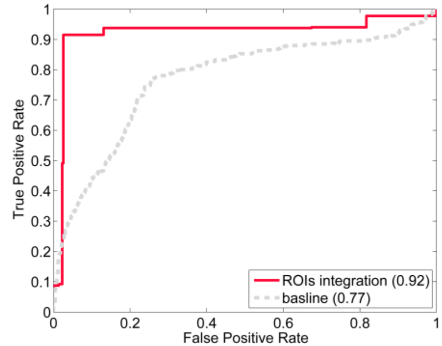


Wheel Region



Linear SVM and top performing descriptors (RGB and Depth) for the instrument panel region and difficult wheel area.

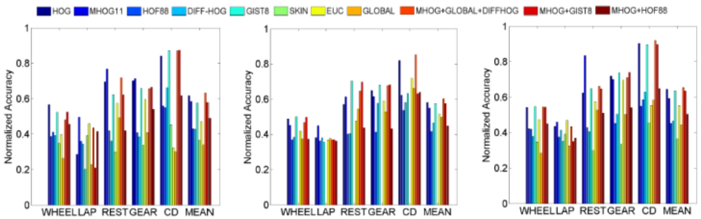
Results – Hand Detection using Integration of Regions



detecting two hands or not in the wheel region

hand activity *as a two class problem* -in the **periphery** regions or the central **wheel** region. The baseline is the top performing descriptor and a first-stage classifier.

Results – No Hand, Hand, and Hand+Object



(a) RGB (b) Depth (c) RGB+Depth

Results – Activity Recognition using Integration of Regions

Results – With Occlusion of Hand-held Objects

Baseline: 35.2%

	WHEEL	LAP	REST	GEAR	CD
WHEEL	.89	.01	.03	.03	.04
LAP	.77	.12	.06		.05
REST	.68		.28	.02	.02
GEAR	.27		.44	.29	
CD	.52		.30	.01	.17

Baseline: HOG-based Deformable Part Model of a mixture over three components. Testing is done at 36 different rotations.

Region Integration (x24 faster)

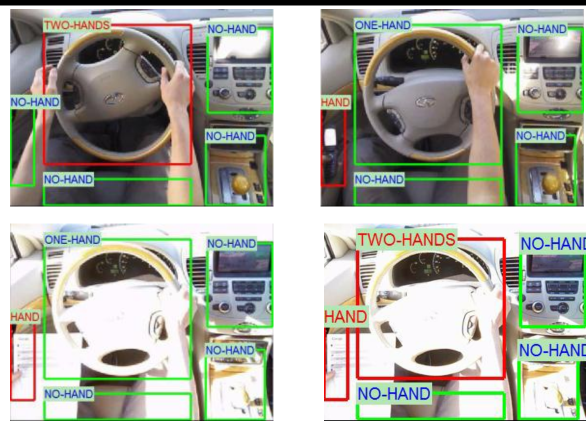
RGB Only: 52.1%

	WHEEL	LAP	REST	GEAR	CD
WHEEL	.84	.04	.04		.08
LAP	.87	.02	.10	.01	
REST	.24		.72		.04
GEAR	.20		.27	.51	.01
CD	.22			.26	.52

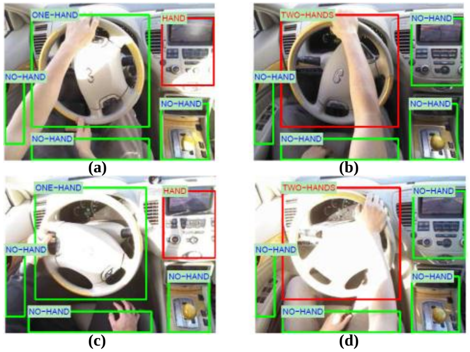
RGB and Depth 69.4%

	WHEEL	LAP	REST	GEAR	CD
WHEEL	.65	.20	.11	.02	.02
LAP	.59	.32	.07	.02	
REST	.01	.01	.97		
GEAR	.10	.03	.26	.58	.03
CD				.04	.95

Results – Activity Recognition using Integration of Regions



Results – Activity Recognition using Integration of Regions

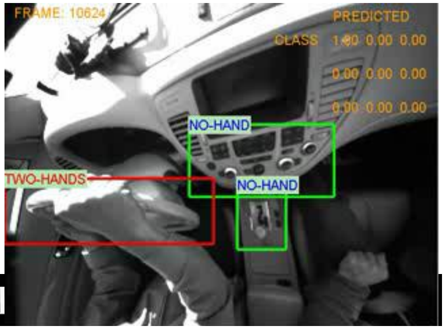


(a) Although the wheel model outputs a prediction of two hands in the wheel region, so does the infotainment due to an illumination artifact. In this case, the integration produces incorrect results since the model learns to give high confidence to the infotainment score.

(b) The lap region produces incorrect classifications due to poor separation in the feature space.

(c) and (d): Illumination produces false positives.

Towards “Real” Real-World – SHRP2



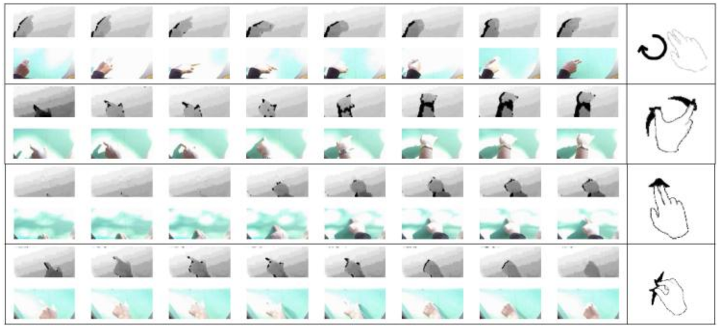
Next Steps

1) Hand and Object Interaction Classes – ***which object?***

2) Efficient extraction and *integration of motion features*

3) Following a hand detection in an ROI *=>*
   User interface through hand gestures

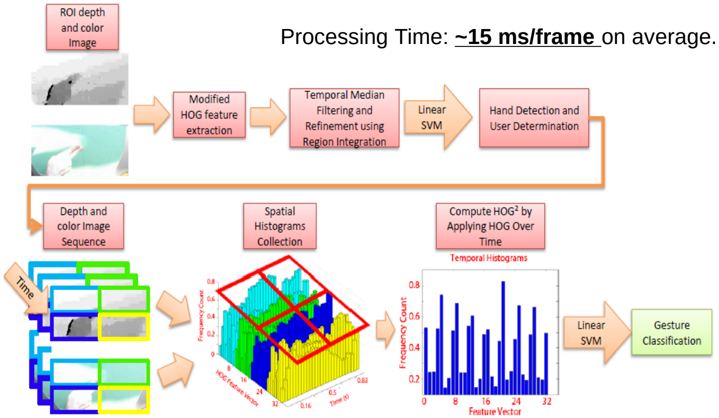
Next Steps – Hand Gesture Recognition in Naturalistic Settings



- **19 Gestures**
- **Illumination Changes**
- **Coarse and fine motion gestures**

Real Time Hand Gesture Recognition – Overall Framework

Processing Time: **~15 ms/frame** on average.



Spatio-Temporal HOG² Descriptor from Color or Depth Images

Modified HOG - Performed at every frame for hand detection. Spatial descriptor.

$$h_t = mHOG(I) = [h^1 \dots h^{M \cdot N}]$$

Spatio-Temporal Feature Extraction (HOG²):

$$\phi(I_1, \dots, I_t) = mHOG\left(\begin{bmatrix} h_1 \\ \vdots \\ h_t \end{bmatrix}\right)$$

Block Normalization of the spatial and temporal histograms:

1) L2-norm: $\phi \rightarrow \phi / \sqrt{\|\phi\|_2^2 + \epsilon}$
2) L2-Hys: L2-norm followed by clipping and renormalization
3) L1-norm
4) L1-sqrt

L2-norm and L1-norm performed best

Extraction Time: ~15 ms/frame on average
State-of-the-art on hand gesture datasets

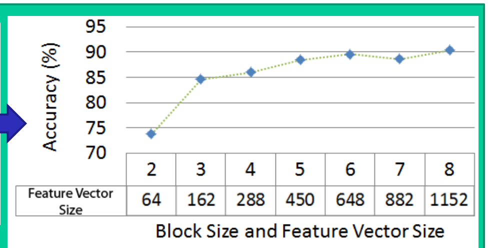
Experimental Evaluation – MSR-Hand Gesture Dataset

- Dataset Statistics:

Depth only,
12 gestures
333 sequences
leave-one-subject-out cross validation.

Method	Accuracy
HOG 3D (Klaser *et al.* [8])	85.23%
HON4D (Oreife and Liu [16])	87.29%
Random Occupancy Patterns (Wang *et al.* [19])	88.5 %
DMM-HOG (Yang *et al.* [21])	89.20%
HON4D + D_{disc} (Oreife and Liu [16])	92.45%
HOG²	**92.64%**

Varying block size in the spatial and temporal stages of the HOG² descriptor for a fixed orientation binning parameter of 9 bins. The figure exhibits the strength of the descriptor even with a *small sized feature set*.



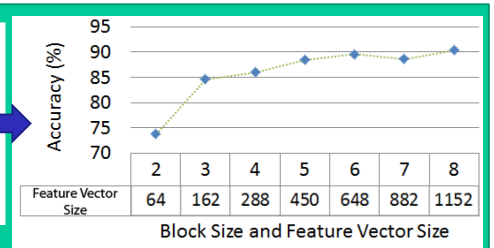
Experimental Evaluation – MSR-Hand Gesture Dataset

- Dataset Statistics:

Depth only,
12 gestures
333 sequences
leave-one-subject-out cross validation.

Method	Accuracy
HOG 3D (Klaser *et al.* [8])	85.23%
HON4D (Oreife and Liu [16])	87.29%
Random Occupancy Patterns (Wang *et al.* [19])	88.5 %
DMM-HOG (Yang *et al.* [21])	89.20%
HON4D + D_{disc} (Oreife and Liu [16])	92.45%
HOG²	**92.64%**

Varying block size in the spatial and temporal stages of the HOG² descriptor for a fixed orientation binning parameter of 9 bins. The figure exhibits the strength of the descriptor even with a *small sized feature set*.



Conclusion

- Studied different *feature extraction methods* for naturalistic hand and hand+object gestures

- Proposed a *cue integration scheme* for constraining the difficult problem of hand detection.

- Exended the spatial features into the temporal domain using HOG^2 , where a modified HOG was applied at every frame, collected into a 2D array, and then applied again.

- Achieved real-time gesture recognition using the state of the art descriptor.