ROBUST REAR-VIEW GROUND SURFACE DETECTION WITH HIDDEN STATE CONDITIONAL RANDOM FIELD AND CONFIDENCE PROPAGATION

Zhiding Yu[⋆] *Wende Zhang*[†] *B. V. K. Vijaya Kumar*[⋆]

[⋆] Department of Electrical and Computer Engineering, Carnegie Mellon University
[†] Electrical Controls and Integration Lab, General Motors Company

ABSTRACT

We address the problem of detecting rear-view (obstacle free) ground surface using a vehicle production camera. This task is considerably more challenging than general front-view road detection, as the associated challenges widely range from low picture quality, fisheye distortion and large objects, to the absence of useful priors such as vanishing points and road structure. Regarding the challenges, we propose a feature that can simultaneously capture local appearance and context information. In addition, the task suffers from strong appearance variations such as shadows and ground markers. Therefore, we propose a novel conditional random field (CRF) model which includes hidden states indicating confident nodes and propagate their confidence to neighboring nodes. We show that our proposed feature and model can jointly achieve robustness against large objects and shadows/markers, showing excellent detection performance under low quality inputs.

Index Terms— ground surface detection, CRF, semantic segmentation, confidence propagation

1 Introduction

Ground surface detection refers to automatically finding obstacle free ground locations in an image. It can benefit future autonomous vehicles and many other intelligent transportation systems. In the past, considerable effort has been made to address the dual problem: road detection [1–4, 16, 24, 25]. However, past works mostly concentrated on front-view structured roads. Few addressed rear-view general ground surface detection, where useful priors such as vanishing points and road structure may not be available. Unlike front-view, pedestrians and other vehicles show up in proximity much more frequently, appearing as large objects. The increased object size further enhances difficulty as image patches from which features are extracted can hardly cover the entire object. The relatively invariant global information therefore becomes less available and classifications are less discriminative. Examples corresponding to these cases are respectively shown in the left and middle columns of Fig. 1.

Another challenge is the misclassification caused by strong shadows and ground markers (See the right column of Fig.1). They look like objects as sharp boundaries generate large responses to feature extractors. One could observe from the examples that such error often occurs at these boundaries. Consequently, a non-robust detection system is likely to exhibit many false alarms in the real world environment.

Fig. 1. Differences between regular front-view (left) and our rear-view scenarios (middle), and false classifications (right). The blue curve indicates pre-defined region of interest (ROI).

Other challenges include the poor picture quality and fisheye distortion. Unlike systems that require high quality sensor inputs (high resolution, vivid saturation and even stereo / depth information), we want our algorithm to be practical enough for the most common sensor: production camera. The method must also be real time to retain its practical value.

The major contributions in this paper are: 1. We propose a feature which jointly captures local appearance and larger-scale context information. 2. We treat coupled superpixels across strongly textured boundaries as mid-level discriminative CRF nodes. 3. We propose a novel hidden state CRF model that can smartly incorporate these discriminative nodes and propagate their label confidence to uncertain nodes. 4. We show our algorithm obtains excellent results with low quality inputs, and can be real-time on general CPUs.

2 Related works

Many past works concentrated on parsing pixel labels with Graphical models [4, 11–15, 26]. Yet their implementations are not possible in real-time[1] on a regular CPU. Some can even be very slow. Predictions of road labels can be subop-

[1] by "real time" we mean around or more than 10 frames per second

timal due to over-fittings that bias towards other categories. Both nearby large objects as well as strong shadows can also cause problems on the detection accuracy.

CRF [29] may be one of the most widely used models for image labeling [10, 30]. Suppose $\mathbf{X}$ denotes the set of observed pixels, $\mathbf{Y}$ the hidden labels of CRF nodes. The joint posterior probability of $\mathbf{Y}$ conditioned on $\mathbf{X}$ is factorized as:

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \prod_i \phi_i(Y_i, \mathbf{X}) \prod_{ij} \psi_{ij}(Y_i, Y_j, \mathbf{X}), \quad (1)$$

and the inference problem is to discriminatively find the label configuration that maximizes the posterior probability:

$$\mathbf{y} = \arg\max_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{X}), \quad (2)$$

where i indexes the CRF nodes corresponding to pixels / superpixels and $j \in N(i)$ indexes nodes neighboring to i. ϕ_i and ψ_{ij} are potential functions satisfying $\phi_i, \psi_{ij} > 0$. $Z(\mathbf{X})$ is the normalization term retaining a probabilistic meaning.

To handle shadows, most works try to compensate shadows with color space properties [5–7] and edge detection [8] [9]. The drawback are: 1. They often rely on ill-conditioned modeling of physical properties of light and cameras. 2. The post-processing can be time consuming[2]. 3. The illumination invariant space may lose discriminative information. We show that with the hidden state CRF model, one can bypass the difficult process of deliberately removing shadows.

3 Proposed Feature and Unary Classification

The algorithm is shown in Fig. 2. We first discuss the proposed feature and the unary (superpixel-wise) classification.

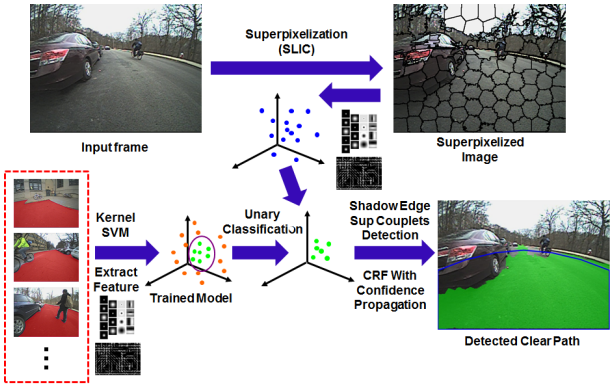

Fig. 2. An illustration of our proposed algorithm flow.

3.1 Superpixelization

We adopt SLIC [17] and extract features from superpixels. Compared with pixel, region provides better semantic information [18, 19] and greatly reduces complexity [20–23].

[2]Take [7] for example. Although the proposed work is linear with the number of pixels, it still requires 0.5s processing time for a 300×400 image.

3.2 Appearance Feature

We model the appearance features using filter banks consisting Gaussians, derivative of Gaussians (DoG) and Laplacian of Gaussians (LoG) at different scales[3].

The appearance feature of each superpixel is then taken as the mean of filter responses within this superpixel. Such feature is highly suitable for modeling ground appearance because it accurately captures both the local color and granular textures at different scales.

3.3 Shape and Contextual Feature

Filter banks alone are not able to cover all types of objects. Due to the low picture quality, large objects often show relatively smooth internal regions similar to ground surface. We need another robust feature to complement filter banks and reject false positives by using more context information.

We extract HOG features [31] from patches centered at the centroid of superpixels. The patches are larger than superpixels to capture more contextual information. In addition, HOG is especially robust and effective for the low quality camera. A HOG spatial pyramid feature is constructed by concatenating the HOG features from co-located patches with different sizes[4]. The final feature is a concatenation of the HOG pyramid and the mean filter bank response.

3.4 Incremental SVM Training

We divide the superpixel features into several subsets and incrementally train an RBF kernel SVM on each subset. Essentially, this is a process of mining hard examples.

4 CRF with Confidence Propagation

We propose a novel CRF model which includes additional hidden states α to indicate confident unary potentials[5]. A graphical representation of our model is shown in Fig. 3. Accordingly, we model the joint posterior probability as:

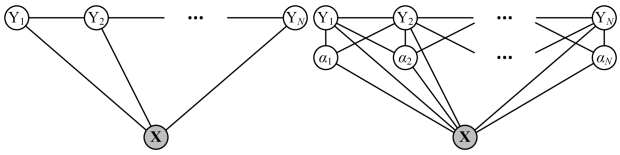

Fig. 3. Graphical representation of conventional CRF model (left) and our proposed hidden state CRF model (right).

[3]We model the filters with different parameters (σ) generated on top of different scales. The corresponding scales we consider are $\sqrt{2}^s$ where scale s ranges from 1 to 3. Under each scale, we generate the gaussian parameters as: $\sigma = s * 2^k, k \in \{0, 1, 2\}$; DoG parameters as: $\sigma = s * 2^k, k \in \{1, 2\}$ and LoG parameters as: $\sigma = s * 2^k, k \in \{0, 1, 2, 3\}$.

[4]A 51×51 and two 101×101 patches are considered at each superpixel location. The first patch is divided into 2×2 cells while the latter two are divided into 3×3 and 4×4 cells.

[5]$\alpha_i \in \{0, 1\}$, $\alpha_i = 1$ indicates that the state of node i is confident. Intuitively, the hidden states of mid-level discriminative superpixels are confident. We will elaborate on how to find these superpixels in Section 4.4.

$$P(\mathbf{Y}, \alpha|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} \prod_i \Phi_i(Y_i, \alpha_i, \mathbf{X}) \prod_{ij} \psi_{ij}(Y_i, Y_j, \alpha_i, \mathbf{X}) \quad (3)$$

where the unary potential and normalization are defined as:

$$\Phi_i(Y_i, \alpha_i, \mathbf{X}) = \phi_i(Y_i, \alpha_i, \mathbf{X}) \delta_i(Y_i, \alpha_i) \varphi_i(\alpha_i, \mathbf{X}) \quad (4)$$

$$Z(\mathbf{X}) = \sum_{\mathbf{Y}, \alpha} \prod_i \Phi_i(Y_i, \alpha_i, \mathbf{X}) \prod_{ij} \psi_{ij}(Y_i, Y_j, \alpha_i, \mathbf{X}) \quad (5)$$

Conventional CRF models often fail upon shadow / marker boundaries because the contrast sensitive Potts model (pairwise potentials) is only good for modeling relatively smooth objects and label can not be correctly propagated across strong edges (See Fig. 1). This requires us turn to more powerful unary classification schemes to compensate the discrimination ability, as intuitively classification works better than smoothness priors in textured regions. The introduced hidden states serve exactly for this purpose.

4.1 Unary potential modelling

In CRF, the unary potential $\Phi_i(Y_i, \alpha_i, \mathbf{X})$ can be regarded as a measure of how likely node i will take on label Y_i and hidden state α_i given the observed image $\mathbf{X}$. Let $\mathbf{f}(.)$ denote the function that maps an arbitrary patch to a feature vector:

$$\mathbf{f} : \mathcal{W}(\mathbf{X}) \mapsto \mathbb{R}^d, \quad (6)$$

where $\mathcal{W}(\mathbf{X}) = \{W_1(\mathbf{X}), ..., W_N(\mathbf{X})\}$ corresponds to superpixels and co-located HOG patches. We model the components of the unary potential Φ as follows:

$$\begin{aligned} \phi_i(Y_i, \alpha_i, \mathbf{X}) &= \exp(P(Y_i|\mathbf{f}(W_i(\mathbf{X}))))^{1-\alpha_i} \\ \delta_i(Y_i, \alpha_i) &= \exp(-S_1(Y_i, \alpha_i)) \\ \varphi_i(\alpha_i, \mathbf{X}) &= \exp(-S_2(\alpha_i, \mathbf{f}'(W_i'(\mathbf{X})))) \end{aligned} \quad (7)$$

where $P(Y_i|\mathbf{f}(W_i(\mathbf{X})))$ can be the output of any discriminative classifiers for superpixel-wise classification, which corresponds to the score of our RBF kernel SVM. Let $Y_i = 1$ indicate that the ground label is true, $S_1(Y_i, \alpha_i)$ is a step function which heavily penalizes discrepancies between Y_i and α_i:

$$S_1(Y_i, \alpha_i) = \begin{cases} \infty & \text{if} \quad Y_i = 0, \alpha_i = 1 \\ 0 & \text{Otherwise} \end{cases} \quad (8)$$

$S_2(\alpha_i, W_i'(\mathbf{X}))$ is also a step function depends on the hypothesis output of the mid-level discriminative node detector:

$$S_2(\alpha_i, \mathbf{f}'(W_i'(\mathbf{X}))) = \begin{cases} \infty & \text{if} \quad \alpha_i \neq H(\mathbf{f}'(W_i'(\mathbf{X}))) \\ 0 & \text{Otherwise} \end{cases} \quad (9)$$

4.2 Pairwise potential modelling

Instead of using the famous contrast sensitive Potts model [28], we propose the relative contrast sensitive Potts model which is more adaptive to local image content:

$$\psi_{ij}(Y_i, Y_j, \alpha_i, \mathbf{X}) = \exp(-\gamma(\mu_{ij}[Y_i \neq Y_j])^{1-\alpha_i}, \quad (10)$$

where μ_{ij} models the coupling strength between two nodes:

$$\mu_{ij} = \frac{1/(||\mathbf{f}_i - \mathbf{f}_j||^2 + \lambda||\mathbf{p}_i - \mathbf{p}_j||^2)}{\sum_{k \in N_i} 1/(||\mathbf{f}_i - \mathbf{f}_k||^2 + \lambda||\mathbf{p}_i - \mathbf{p}_k||^2)}, \quad (11)$$

and $\mathbf{p}_i$ denotes the image plane location of the superpixel centroid. $\mathbf{f}_i \triangleq \mathbf{f}(Wsup_i(\mathbf{X}))$ is the mean filter bank response of suerpixel (node) i. λ is the parameter which decides the weight of spatial information and is empirically set to 0.5.

4.3 Inference

We use iterated conditional modes (ICM) to greedily infer the approximate label configuration which maximizes the posterior probability. Taking the log of $P(\mathbf{Y}, \alpha|\mathbf{X})$, we have:

$$\begin{aligned} \log(P(\mathbf{Y}, \alpha|\mathbf{X})) \propto & \sum_i ((1-\alpha_i)P_i - S_{1,i} - S_{2,i}) \\ & + \gamma \sum_i \sum_j (1-\alpha_i)\mu_{ij}[Y_i \neq Y_j]. \end{aligned} \quad (12)$$

One can infer the labels with a two-step maximization. The first step is to independently maximize with respect to α_i, $\forall i$:

$$\alpha_i = \arg\max_{\alpha_i} \log(P(\mathbf{Y}, \alpha|\mathbf{X})) = \arg\max_{\alpha_i} \sum_i -S_{2,i} \quad (13)$$

Note that we simplify the maximization action to maximizing $\sum_i -S_{2,i}$ because this is the only way to avoid minus infinity. The second step is to iteratively and independently maximize with respect to each label configuration Y_i, $\forall i$:

$$\begin{cases} Y_i = 1 & \text{if } \alpha_i = 1 \\ Y_i^{(k+1)} = \arg\max_{Y_i} \sum_i \alpha_i P_i & \\ \quad + \gamma \sum_i \sum_j (1-\alpha_i)\mu_{ij}[Y_i \neq Y_j^{(k)}] & \text{if } \alpha_i = 0 \end{cases} \quad (14)$$

4.4 Hidden state via shadow/marker edge detection

Ground shadow edges are good discriminative mid-level cues that can generate trustable unary predictions. We can incorporate them into our hidden state CRF and improve the labeling result significantly with limited additional computation cost.

A single superpixel can be much less discriminative than a coupled superpixel pair across shadow edge since the latter is relatively texture-rich. Therefore we treat the coupled ground superpixels along a shadow edge as confident nodes in the CRF. We find out all such couplets and simply concatenate their filter bank features.[6] We then train an RBF kernel SVM to detect such couplets in a test image. The detection setup for ground markers is exactly same. Here, the SVM for couplet detection corresponds to the hypothesis $H(\mathbf{f}'(W_i'(\mathbf{X})))$ in Eq. (9). The hidden states α_i, α_j of any detected couplet (i, j) will be set as confident, based on the first step of our inference. And because we are very confident that the fired couplets are ground surfaces,[7] the labels Y_i of such couplets will be fixed as positive in inference. (See Eq. (14).)

[6] There is a fixed order for couplets, i.e., always choose the shadowed ground superpixel as the first one in feature concatenation.

[7] By definition both coupled superpixels in true couplets must be ground.

5 Experiment

We conduct ground surface detection experiments on self collected production camera data. The Alpine HCE-C115 analog rear-view camera is used to collect videos.

Fig. 4 gives examples of our collected dataset and the pre-defined ROIs. We labeled over 1500 images in which 593 images are training set while the rest are generated from two different sequences. One is mildly shadowed and the other is more challenging with both large objects and strong shadows.

Fig. 4. Some examples of the dataset and ROIs.

Fig. 5 illustrates some of the intermediate results of detected coupled shadow edge superpixels. One could see that such method works exactly as expected.

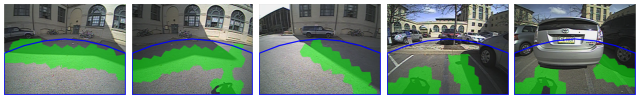

Fig. 5. Illustration of shadow edge detection using coupled cross-boundary superpixels.

We compare our feature extraction method with the Gabor filter bank used in [2] on images containing nearby objects. We use exactly the same feature extraction scheme as [2] except the perspective rectangular patches are replaced with superpixels. Fig. 6 illustrates superpixel-wise classifications.

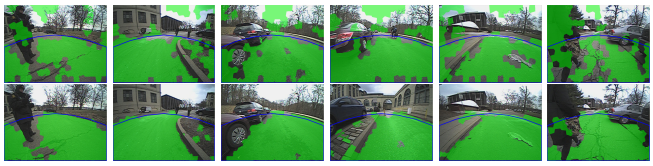

Fig. 6. Qualitative comparison between [2] (top) and the proposed (bottom) method on features.

Experiment shows our method generates much better results than Gabor filters used in [2] under low quality images. The method used in [2] gives a lot of false positives in relatively smooth regions of nearby objects. Our feature on average gives 88.7% of accuracy while [2] only gives 74.4%[8].

On the mildly shadowed sequence the estimated accuracy of our complete method reaches over 90% for all ROIs. Some qualitative results are shown in Column 1 and 2 of Fig. 7 and a complete video is available at [32]. The final labeling accuracy is listed in Table 1. Intuitively, one is more interested in the detection accuracy of nearby ground surface than the accuracy of distant ground. In this experiment, as the region of interest gets closer, the detection accuracy increases.

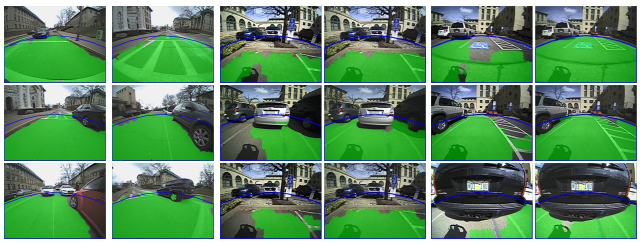

Fig. 7. Some qualitative results on the two test sequences.

Table 1. Accuracy on mildly shadowed sequence.

	ROI1	ROI2	ROI3
Accuracy	93.77%	94.83%	95.26%

We finally conduct experiments on the more challenging video sequence. We use our feature with conventional CRF as the baseline. Fig. 7 shows the results of the CRF baseline (Column 3 and 5) and the hidden state CRF (Column 4 and 6). Kindly refer to [33] for the complete video result. Our method is able to overcome the shadow / marker problem and correctly propagates ground label into shadowed regions. The method is also very robust to large objects, by correctly distinguishing them as non clear path. In addition to baseline we also compare to [2] and [6]. Quantitative results are listed in Table 2, showing better performance of our method.

Table 2. Accuracy on strongly shadowed sequence.

Accuracy	ROI1	ROI2	ROI3
CRF Baseline	89.03%	91.01%	91.52%
Hidden State CRF	$\mathbf{92.92\%}$	$\mathbf{94.15\%}$	$\mathbf{94.54\%}$
Method of [2]	84.01%	84.42%	85.30%
Method of [6]	87.17%	87.91%	89.39%

6 Conclusions

We proposed a novel hidden state CRF model and used this model for ground surface detection. The method generates good clear path detection results under low image quality while being robust to large objects and strong shadows. A real-time implementation of this work has also been done in C++ which processes nearly 10 frames per second on an i7 3940XM CPU, showing great practical value of this method.

[8] These are the results of basic unary (superpixel-wise) classification.

7 References

[1] Q. Wu, W. Zhang, B. V. K. Vijaya Kumar, "Example-based clear path detection assisted by vanishing point estimation," *ICRA*, 2011.

[2] Q. Wu, W. Zhang, T. Chen, B. V. K. Vijaya Kumar, "Camera-based clear path detection," *ICASSP* 2010.

[3] O. Miksik, P. Petyovsky, L. Zalud, P. Jura, "Robust Detection of Shady and Highlighted Roads for Monocular Camera Based Navigation of UGV," *ICRA* 2011.

[4] O. Miksik, D. Munoz, J. A. Bagnell, M. Hebert, "Efficient Temporal Consistency for Streaming Video Scene Analysis," *ICRA* 2013.

[5] C. Lu, M.S. Drew, G.D. Finlayson, "On the Removal of Shadows From Images," *IEEE Trans. PAMI*, vol. 28, pp. 59-68, Jan. 2006.

[6] Q. Wu, W. Zhang, B.V.K. Vajaya Kumar, "Strong shadow removal via patch-based shadow edge detection," *ICRA*, 2012.

[7] M.S. Drew, C. Lu, G.D. Finlayson, "Entropy Minimization for Shadow Removal," *IJCV*, pp. 35-57, 2007.

[8] J. F. Lalonde, A. A. Efros, S. G. Narasimhan, "Detecting Ground Shadows in Outdoor Consumer Photographs," *ECCV*, 2010

[9] A. Gijsenij, T. Gevers, "Shadow edge detection using geometric and photometric features," *ICIP*, 2009

[10] J. Shotton, "TextonBoost for image understanding, Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *IJCV*, vol. 81, pp. 2-23, 2009.

[11] S. Gould, R. Fulton and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," *ICCV*, 2009.

[12] L. Ladicky, C. Russell, P. Kohli and P. Torr, "Associative hierarchical crfs for object class image segmentation," *ICCV*, 2009.

[13] J. Lim, P. Arbelaez, C. Gu and J. Malik, "Context by region ancestry," *ICCV* 2009.

[14] L. Zhang and Q. Ji, "Image segmentation with a unified graphical model," *IEEE Trans. PAMI*, vol. 32, pp. 1406-1425, 2010.

[15] R. Mottaghi and S. FidlerAnalyzing, "Semantic Segmentation Using Human-Machine Hybrid CRFs," *CVPR*, 2013.

[16] Z. He et al., "Robust Road Detection from A Single Image Using Road Shape Prior," *ICIP*, 2013.

[17] R. Achanta et al., "SLIC Superpixels Compared to State-of-the-art Superpixel Methods," *IEEE Trans. PAMI*, vol. 34, pp. 2274-2282, 2012.

[18] Y. Zhou, L. Li and H. Zhang, "Adaptive Learning of Region-based pLSA Model for Total Scene Annotation," *arXiv*:1311.5590, 2013.

[19] Y. Zhou et al., "Region-based high-level semantics extraction with CEDD," *IEEE Int. Conf. on Network Infrastructure and Digital Content*, 2010.

[20] Z. Yu et al., "Nonparametric Density Estimation on A Graph: Learning Framework, Fast Approximation and Application in Image Segmentation, *CVPR*, 2011.

[21] Z. Yu et al., "Automatic Object Segmentation from Large Scale 3D Urban Point Clouds through Manifold Embedded Mode Seeking," *ACM-MM*, 2011.

[22] Z. Yu, A. Li, O. C. Au and C. Xu, "Bag of Textons for Image Segmentation via Soft Clustering and Convex Shift," *CVPR*, 2012.

[23] W. Liu, Z. Yu and D. Meng, "Joint Recognition / Segmentation with Cascaded Multi-level Feature Classification and Confidence Propagation," *ICME*, 2013.

[24] J. Crisman, C. Thorpe, "Unscarf, a color vision system for the detection of unstructured roads," *ICRA*, 1991.

[25] J. Crisman, C. Thorpe, "Scarf: A color vision system that tracks roads and intersections," *IEEE Trans. Robotics and Automation*, vol. 9, pp. 49-58, 1993.

[26] D. Munoz, J. A. Bagnell, M. Hebert, "Stacked Hierarchical Labeling," *ECCV*, 2010.

[27] R. Potts, "Some Generalized Order-Disorder Transformation," *Proc. Cambridge Philosophical Soc.*, vol. 48, pp. 106-109, 1952.

[28] Y. Boykov G. Funka-Lea, "Graph Cuts and Efficient N-D Image Segmentation," *IJCV*, pp. 109-131, 2006.

[29] J. Lafferty, A. McCallum, F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *ICML*, 2001

[30] S. Kumar and M. Hebert, "Discriminative random fields," *IJCV*, vol. 68, pp. 179-201, 2006.

[31] O. Ludwig, D. Delgado, V. Goncalves and U. Nunes, "Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection," *International IEEE Conference On Intelligent Transportation Systems*, 2009.

[32] http://youtu.be/RqZhj6JiLaY

[33] http://youtu.be/kHpmLmZR5vw