Latent Hierarchical Part Based Models for Road Scene Understanding

Suhas Kashetty Venkateshkumar*
Continental AG
Lindau, Germany
Suhas.Kashetty.Venkateshkumar@continental-corporation.com

Muralikrishna Sridhar
Continental AG
Lindau, Germany
Muralikrishna.Sridhar@continental-corporation.com

Patrick Ott†
School of Computing
University of Leeds
p.ott@leeds.ac.uk

Abstract

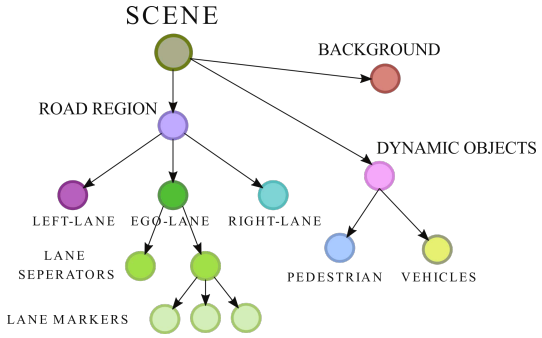
Road scenes can be naturally interpreted in terms of a hierarchical structure consisting of parts and sub-parts, which captures different degrees of abstraction at different levels of the hierarchy. We introduce Latent Hierarchical Part based Models (LHPMs), which provide a promising framework for interpreting an image using a tree structure, in the case when the root filter for non-leaf nodes may not be available. While HPMs have been developed in the context of object detection and pose estimation, their application to scene understanding is restricted, due to the requirement of having root filters for non-leaf nodes. In this work, we propose a generalization of HPMs that dispenses with the need for having root filters for non-leaf nodes, by treating them as latent variables within a Dynamic Programming based optimization scheme. We experimentally demonstrate the importance of LHPMs for road scene understanding on Continental and KITTI datasets respectively. We find that the hierarchical interpretation leads to intuitive scene descriptions, that is central for autonomous driving.

1. Introduction

We address the task of understanding a scene in terms of a Latent Hierarchical Part based Model (LHPM). Such models provide a way of interpreting the whole scene in terms of the geometric relationships to its respective parts. Each of these parts can in turn be recursively expressed in terms of geometric relationships to their respective sub-parts. As illustrated in Figure 1, the visual world naturally lends itself to a tree based hierarchical representation that captures different degrees of abstractions at different levels. The parts of the hierarchy tend to have a natural geometric relationship with each other.

The application of part based models to scene understanding [4, 22] have so far been restricted to non-hierarchical (star-shaped) models. Hierarchical extensions to part based models, in the form of Deformable Part based Models (DPM) [9, 8, 21, 1], have been largely applied to object detection and human pose estimation [30, 1, 31] respectively. However, these approaches rely on having root filters (detectors) corresponding to higher levels in the hierarchy. This poses a problem for extending hierarchical part

(a) Road scene with scene description



(b) Hierarchical decomposition of the scene in (a)

Figure 1. (a) A typical road scene with multiple lanes, dynamic objects and static background. (b) Our approach interprets a road scene using a Latent Hierarchical Part based Model (LHPM), which dispenses the need for having root filters for non-leaf nodes in the hierarchy.

*This research work was carried out at Continental and submitted as master thesis at the Ingolstadt University of Applied Science, Germany.

†Patrick Ott is an Honorary Research Fellow of the University of Leeds. This research was undertaken during his time at the Continental.

based model for scene understanding as it is generally not feasible to have root filters for many regions in a scene, *e.g.* for a road region.

Our contribution is to extend HPMs in a way that dispenses with the requirement for having root filters for the non-leaf nodes of the hierarchy. Our extension involves treating the non-leaf nodes as *latent* and inferring them in a recursive manner within a optimization scheme based on Dynamic Programming (DP). Our approach extends regular HPMs [9, 8, 21] to scene understanding, where root filters are often an impractical idiosyncrasy.

Our experiments demonstrate the importance of LHPMs for road scene understanding on two datasets capturing diverse road scenes. We show that encoding geometric relations at multiple levels results in a significant reduction of errors. Moreover, we also show that the proposed LHPMs deliver an intuitive interpretation of the scene. We demonstrate the relevance of an interpretation specifically for autonomous driving.

We organize the paper in the following manner. The following section describes the related work. Section 3 details the formulation of LHPMs and presents the challenge due to missing root filters for non-leaf nodes. Section 4 describes our contribution towards incorporating latent estimation of non-leaf nodes. Section 5 discusses the application of our approach to road scene understanding. Section 6 is dedicated to evaluation and experimental analysis. Section 7 summarizes our main contributions and outlines future research directions.

2. Related Work

Much of the research in scene understanding has approached this problem from the perspective of segmentation of pixels [6, 2] into semantically meaningful regions. A Conditional Random Field (CRF) [18] is typically used to model the appearance around a homogeneous region (*e.g.* pixels, superpixels), combined with the spatial relations between neighboring regions. CRFs have been extended by Ladickỳ *et al*. [16, 23] to model hierarchical relations between regions. Wojek *et al*. [28] proposed a novel approach based on conditional random field to jointly perform object detection and scene labeling. However, learning the optimal components of the model, especially with several layers in the hierarchy, is practically intractable due to complex dependencies. To address such complexities, Munoz *et al*. [20] proposed a stacked hierarchical inference on a graphical model, which breaks the complex inference process into a hierarchical series of sub-problems. However, segmentation based techniques tend to create a noisy segmentation in real road scenes, as pointed out by Miksik *et al*. [19], who tried to resolve this issue using temporal smoothing techniques (*e.g.* optical flow).

In contrast to the above segmentation-based approaches, part based models offer a principled way of "searching" for the right configuration of parts that can make up a whole scene.

Closely related to our work is from authors Topfer *et al*. [26], who apply a hierarchical part based model for interpreting a road scene. Their basic building block is the detection of a patch, which is defined in terms of a pair of parallel lane-marking features. They recursively build lanes by stacking these patches together explicitly in their graphical representation. Spatial relations between patches and between lanes, are represented by continuous distributions and non-parametric belief propagation is used to perform inference in this framework. Their approach relies on the presence of a sequence of patches that would constitute a lane. In contrast to this work, our approach avoids linearly stacking up patches to build lanes. The novelty of our approach lies in treating the positions of lane separators, lanes and road region as latent variables and estimating these in our inference scheme.

Pandey & Lazebnik [22] apply part based model for scene understanding. However, the application of part based models to scene understanding has so far been restricted to star-shape models, where the root node is also detected using a dedicated filter.

The significance of hierarchy for part based models has been demonstrated in the case of object detection [11, 8, 21, 1] and human pose estimation [30, 1, 31] respectively. However, their extension to scene understanding remains restricted since these approaches, again, rely on having root filter(s) (*i.e.* detectors) corresponding to higher levels in the hierarchy, and it is generally not feasible to have root filters for many regions in a scene (*e.g.* for a road region).

Our contribution is to generalize HPMs to road scene understanding, where root filters are often impractical. We show that a more complex *hierarchical structure* imposes strong constraints that enables us to efficiently search for the right configuration of parts that can make up a whole scene, despite errors arising from object detectors at the bottom level. Moreover, by dispensing with the need of having root filters for the non-leaf nodes, we are able to extend HPMs to road scene understanding.

3. Formulation

Given an image $\mathcal{I}$ of a road scene, we would like to interpret this image in terms of a hierarchical part based model, such as the one illustrated in Figure 1. Such a model aptly represents a scene in terms of the relationship between parts and their respective sub-parts. The tree has a root node in form of a global scene node, which captures the static aspect as shown in Figure 1. This root node is decomposed in terms of its respective parts, each of which in-turn are split into further parts at the next level of the tree structure.

More formally, a node in the tree is given by v and the

set of all nodes in the tree are given by $\mathcal{V}$. The leaf nodes are $\mathcal{V}_f \subset \mathcal{V}$. Any node v except the root node has an unique parent given by $pa(v)$. Similarly, any node v, except the leaf nodes, has a set of children given by the set $ch(v)$.

The random variable l_v is a hypothesis for the location of the node v in the image $\mathcal{I}$. For any particular non-leaf node v with position l_v, we define a configuration $L_{ch(v)}$ as a set of children nodes $L_{ch(v)} = \{..., l_{v'}, ...\}$. The random variable $\mathcal{L} = \{..., l_v, ...\}$ represents one particular hypothesized scene configuration consisting of locations of all the nodes in the tree.

Our objective is to infer the optimal scene configuration $\hat{\mathcal{L}}$ that maximizes the posterior probability distribution $\mathrm{P}(\mathcal{L}|\mathcal{I}, \Theta)$, given $\Theta = \{..., \theta_v, ...\}$, the model set for all the nodes in the tree.

$$\hat{\mathcal{L}} = \arg\max_{\mathcal{L}} \mathrm{P}\big(\mathcal{L}|\mathcal{I}, \Theta\big) \tag{1}$$

The hierarchy imposes natural conditional independence assumptions, using which we can factorize the posterior as

$$\mathrm{P}(\mathcal{L}|\mathcal{I}, \Theta) \propto \prod_{v \in \mathcal{V}_f} \mathrm{P}\big(l_v|\mathcal{I}, \Theta\big) \prod_{v \in \mathcal{V} \setminus \mathcal{V}_f} \mathrm{P}\big(l_v|L_{ch(v)}, \Theta\big) \tag{2}$$

This probability first of all implies that the location l_v of any particular leaf node $v \in \mathcal{V}_f$ is directly constrained by the appearance probability $\mathrm{P}\big(l_v|\mathcal{I}, \Theta\big)$ of placing a window at location l_v over the image $\mathcal{I}$, provided by the object detector. Secondly, the location l_v of any non-leaf node v is indirectly constrained by the locations $L_{ch(v)} = \{l_{v'} : v' \in ch(v)\}$ of their respective children nodes $ch(v)$ according to the spatial probability $\mathrm{P}\big(l_v|L_{ch(v)}, \Theta\big)$.

In a nutshell, Equation 2 expresses our key objective *i.e.* the *optimal* configuration $\hat{\mathcal{L}}$ is the one that maximizes the posterior $\mathrm{P}\big(\mathcal{L}|\mathcal{I}, \Theta\big)$.

4. Optimization

In order to optimize Equation 2 in an efficient manner, we use dynamic programming (DP), which has been considered a natural choice for tree structures [9, 8]. The DP based optimization is expressed as follows.

$$\mathcal{B}(l_v) \quad = \quad \max_{l_v} \Big(\mathrm{P}\big(l_v|L_{ch(v)}\big) \prod_{v' \in ch(v)} \mathcal{B}(l_{v'})\Big) \tag{3}$$

The DP formulation given above expresses the probability $\mathcal{B}(l_v)$ of the optimal location l_v of a node v, which is recursively expressed in terms of (i) the spatial probability $\mathrm{P}\big(l_v|L_{ch(v)}\big)$; (ii) the accumulated optimal production probabilities of all its children given by the product $\prod_{v' \in ch(v)} \mathcal{B}(l_{v'})$. For the leaf nodes, the production probability is given by the appearance probability $\mathrm{P}\big(l_v|\mathcal{I}, \Theta\big)$ or more simply $\mathrm{P}(l_v)$.

The DP formulation assumes that the optimal position $\hat{l}_v$ for each node v requires the optimal configuration $\hat{L}_{ch(v)}$ of its children nodes to be already found. Thus, the inference procedure starts from the leaf nodes and proceeds upwards recursively until the root node is reached. When the root node of the hierarchy is reached, we can read off exactly which configurations led to this global maximum, by reconstructing which predecessor states led to the best result. Arriving back at the leaf nodes, our solution is complete.

In order to find the optimal configuration for any node v, we have to generate multiple configurations $L_{ch(v)}$ for all its children nodes and score each of them as follows.

$$\mathcal{B}(l_v) = \max_{l_v} \prod_{l_{v'}} \Big(\mathrm{P}\big(l_{v'}|l_v\big) \prod_{v' \in ch(v)} \mathcal{B}(l_{v'})\Big) \tag{4}$$

In previous works on applying HPMs to object detection [9, 8, 21, 1] and human pose estimation [30, 1, 31] respectively, the position l_v of the root node, required to compute the probability $\mathrm{P}\big(l_{v'}|l_v\big)$ is generally provided by the use of a root filter (i.e. appearance probability given by an object detector). Such a setup can significantly help in keeping the search space in the aforementioned energy surface more tractable. However, it is not feasible to have a root filter for scene entities that are hard to detect using a sliding window based detector (*e.g.* road regions). Therefore, we here treat the root node as a global scene node.

We compute the expected position $E(l_v|L_{ch(v)})$ of a latent parent node v, given a configuration $L_{ch(v)}$ of its children nodes as follows.

$$E(l_v|L_{ch(v)}) = \sum_{l_{v'} \in L_{ch(v)}} E(l_v|l_{v'})\mathrm{P}(l_{v'}) \tag{5}$$

Here, $E(l_v|l_{v'})$ is the expected position l_v of a latent parent node v with respect to a single child node $l_{v'} \in L_{ch(v)}$. This expectation is in turn computed by taking the expectation of l_v given the conditional probability distribution $\mathrm{P}(l_v|l_{v'}, \theta_v)$, as follows.

$$E(l_v|l_{v'}) = \int l_v \mathrm{P}(l_v|l_{v'}, \theta_v) dl_v \tag{6}$$

The distribution $\mathrm{P}(l_v|l_{v'}, \theta_v)$ given in Equation 6 specifies the probability of the position of l_v given a child node's position $l_{v'}$ using an appropriate probability distribution with parameters θ_v. We rewrite Equation 4 using the expected position l_v which is computed using Equation 5 and 6 as follows

$$\mathcal{B}(l_v) = \max_{l_v} \prod_{l_{v'}} \Big(\mathrm{P}\big(l_{v'}|E(l_v|L_{ch(v)}), \theta_v\big) \prod_{v' \in L_{ch(v)}} \mathcal{B}(l_{v'})\Big) \tag{7}$$

The inference procedure proceeds upwards in the tree by iteratively using Equation 7 to score possible locations for

each node in the tree, until the scene node is reached. At this point, the optimal configuration for the entire tree is the set of configurations for each node that maximizes Equation 7.

5. Application to Road Scene Understanding

In this section, we describe the application of LHPMs to the proposed hierarchical interpretation of road region as shown in Figure 1. For the sake of illustration we consider a part of this hierarchy that corresponds to the road region, with four levels corresponding to lane markers, lane separators, lanes and the road region respectively. The road region is especially interesting because the use of root filters to detect higher level entities (such as lane separators and lanes) is in-feasible using sliding window based object detectors. The proposed LHPMs addresses this issue by treating the higher level entities as *latent* nodes in the inference procedure as described in Section 4.

We now describe the details of how the latent position $\hat{l}_v$ of a parent node v is estimated given a certain configuration $L_{ch(v)} = \{..., l_{v'}, ...\}$ using two steps. Firstly, each child node $v' \in ch(v)$ given its respective location $l_{v'}$, projects the expected position $E(l_v|l_{v'})$ of the reference node v using the model for the node θ_v as given by Equation 6. In our application, we use a bi-variant normal distribution to model the spatial relationship a parent and child node.

For level 1 in the hierarchy depicted in Figure 2 (a), the expected position of a lane separator is projected by the lane markers in level 0, as shown using the black markers in Figure 2 (b). Similarly, the expected positions of the lanes and the road region are each projected by the lane separators and lanes respectively for the next two levels of the hierarchy, This is illustrated in Figure 2 (e) for level 2 and Figure 2 (h) for level 3 respectively.

In the second step, the expected position $E(l_v|L_{ch(v)})$ of l_v given a configuration $L_{ch(v)}$ of its child nodes is computed using Equation 5 and is illustrated using an orange marker for level-1 (lane separators) in Figure 2 (b), an orange line for level-2 (lanes) in Figure 2 (e) and level-3 (road region) in Figure 2 (h) respectively.

Having inferred the expected position $E(l_v|L_{ch(v)})$ of the parent node v, we use the learned model θ_v to compute the spatial probability $\mathrm{P}\big(l_{v'}|E(l_v|L_{ch(v)}), \theta_v\big)$ that appears in Equation 7. The bi-variate distributions given by parameters θ_v are illustrated for all three levels in Figure 2 (c) (lane separators), Figure 2 (f) (lanes) and Figure 2 (h) (road region).

6. Experiments

We report our results on two datasets. The first dataset was collected by Continental AG, depicting a variety of traffic scenarios with different types of lane configurations. This dataset contains multi-lane annotations, enabling us to perform more detailed lane-based evaluation for validating our approach. This dataset is divided into 150 images for training and 150 images for testing.

The second dataset is the KITTI Vision Dataset [10]. This dataset contains a set of 95 images, which were extracted from video footage of driving around urban roads, and was manually annotated with binary labels specifying whether each pixel is drivable road or otherwise.

In this work, we have used the Continental training dataset for training our object detectors and learning the parameters of our LHPMs. We report our results on the Continental test set, comprising of 150 images. We also use the *same models* to report our performance on the KITTI dataset. Our results given in the next subsection show the robustness of using models trained on the Continental dataset in extrapolating to the KITTI dataset, without any extra training.

We first ran a sliding window based object detector for the *lane markers*, *pedestrians* and *cars* respectively. We used Histograms of Oriented Gradient (HOG) feature descriptors [5] and a *linear* SVM classifier [27]. We deliberately kept a low detection and non-maximum suppression threshold, generating noisy detections as shown in the first column of Figure 4 and Figure 5. We then feed these detections to the proposed LHPM. We obtain results at different levels of the hierarchy, namely person and car for moving objects, and a configuration of lane markers, lane separators and lanes respectively. We limit the lane detection upto 3 lanes (*i.e.* left-lane, ego-lane and right-lane) in this work. Finally, we use a standard $\mathcal{F}1$ computed based on PASCAL VOC guidelines [7], as a scoring criterion to evaluate the performance of the proposed LHPM.

6.1. Quantitative Evaluation

In this section, we present a quantitative evaluation of our approach. The results show that the performance LHPMs produce promising performance on both Continental and KITTI dataset.

More significantly, we would like to assess the role of LHPMs for hierarchical scene interpretation. We first present a comparison of the performance of detecting the ego lane at different levels of the tree based representation. Figure 3 (c) for the KITTI dataset and Figure 3(a) for the Continental dataset, represents the performance of detecting the ego-lane at various levels. In these plots, the $\mathcal{F}$-measure is plotted on the y-axis and a_o area of overlap (expressed in percentage) is plotted on the x-axis. We observe from these two plots that the performance at level-1 is comparatively lower to level-2 and level-3 for all overlap values.

Our second comparison consists of evaluating the performance difference between level-2 and level-3, in detecting all the three lanes (*i.e.* ego-lane, left-lane, right-lane) and plotting the resulting $\mathcal{F}$-measure *vs* a_o – see Figure 3 (b).

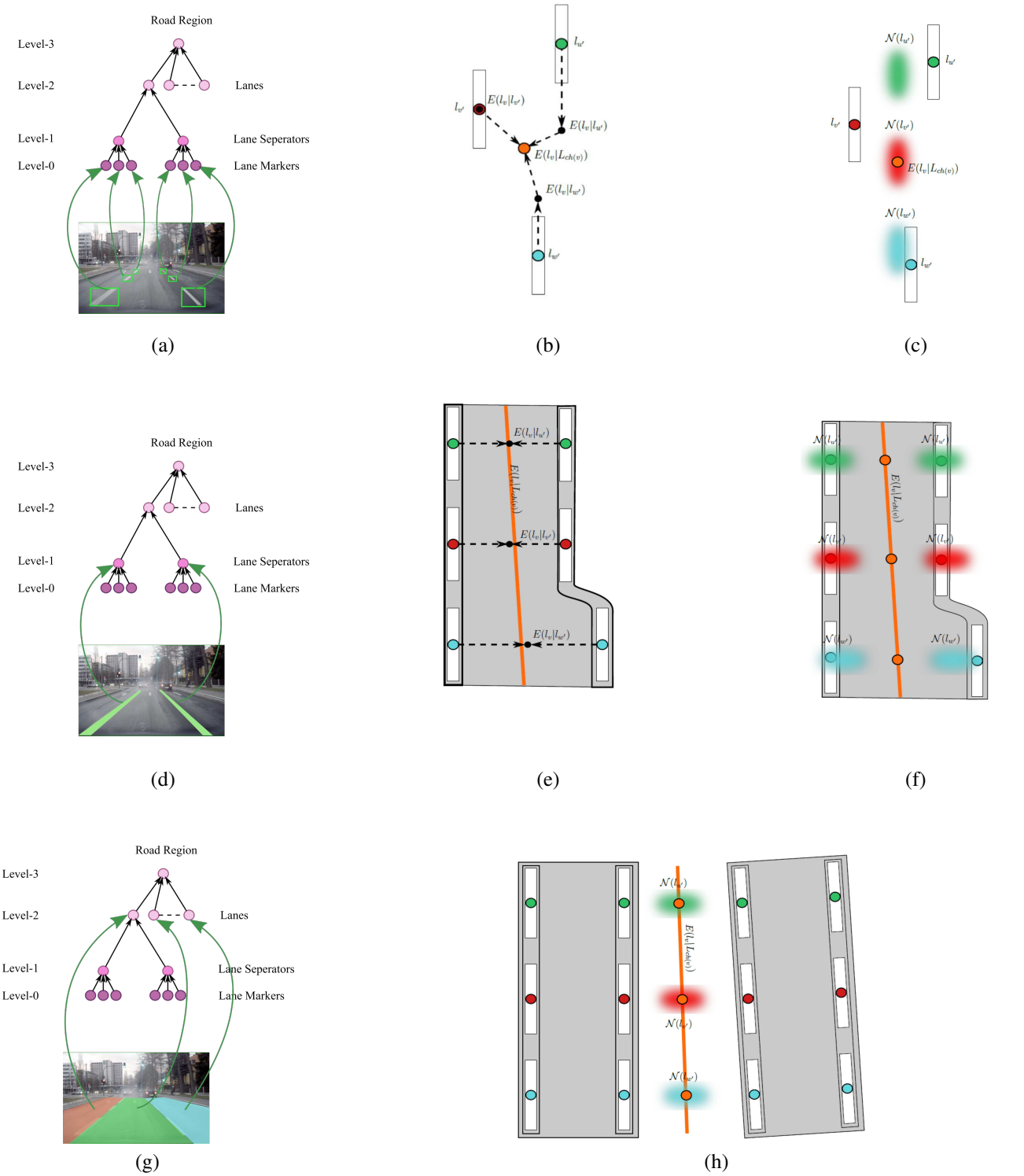


Figure 2. This figure is used to describe the computation of the latent node's position for all three levels of the road region hierarchy. The details are described in the main text.

We observe a similar result as before: the performance at level-3 is always better than that at level-2 as one would suspect as more scene information is taken into account. (Note: Level-1 is not taken into consideration in this comparison, since at level-1 the objective is to detect lane separator and has no information of how a lane/road-region looks like.

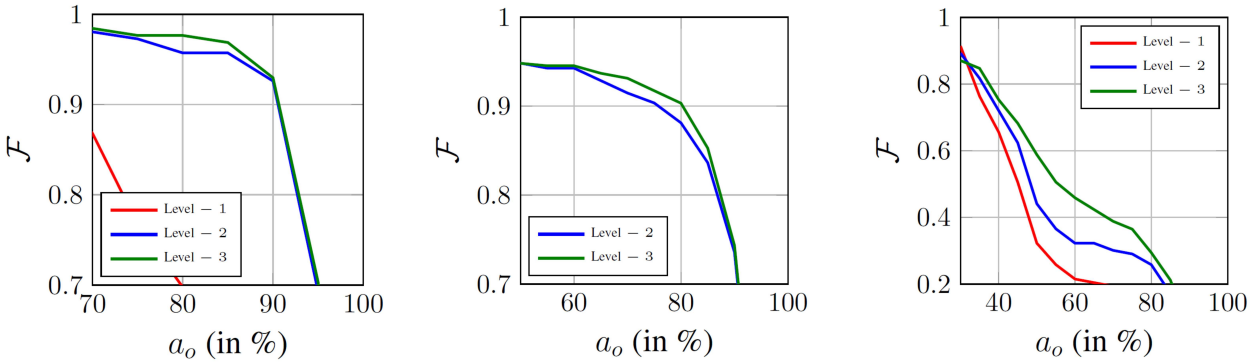


Figure 3. The above plots report the performance in terms of the $\mathcal{F}$-measure for different values of overlap a_o. Sub-Figure (a) compares the performance of detection performance for the ego lane at levels 1, 2 and 3 respectively, on the Continental dataset. Sub-Figure (b) depicts the detection performance for left, ego and right lanes at levels 2 and 3 respectively, on the Continental dataset. Sub-Figure (c) depicts the detection performance for the ego lane at levels 1, 2 and 3 respectively, on the KITTI dataset.

UM - Perspective space				
Method	F_{max}	**Prec.**	**Recall**	**Acc**
BL [10]	88.9	87.3	90.6	95.3
SPRAY [15]	88.3	90.7	86.0	95.2
LHPM	87.0	60.5	75.2	94.0

Table 1. Results [%] of pixel-based ego-lane evaluation in perspective space, on the KITTI dataset for Urban Marked (UM) road scenes.

Thus, a comparison with level-2 and level-3 would be unfair.)

These results support the following conclusion: an LHPM-based approach with multiple levels, leading to a road scene understanding, can effectively reduce errors, given that more information is included at multiple levels in a structured manner.

6.2. Qualitative Evaluation

For a qualitative analysis we direct the readers attention to Figure 5 for the Continental dataset and Figure 4 for the KITTI dataset respectively.

These two figures depict detection results on images as represented by rows. The first column shows lane marker detections at a low detection threshold (color coding: red to green in decreasing order of detection score). The second and third columns show detection results at level-2 and level-3 respectively (color coding: (i) red represents left-lane; (ii) green represents ego-lane and (iii) blue represents right-lane). Detections of the pedestrians and vehicles are also shown in order to infer the relative locations of these moving objects with respect to these different lanes.

In Figure 5 (a,c,d), despite noisy detections, both Level 2 and 3, have been able to predict all three lanes. However, in Figure 5 (b), there is a significant amount of false positives on the right lane. Although these false positives are of low scores (bright green boxes in first image from left), yet their spatial arrangement yields a better lane hypothesis when compared to the spatial arrangement of the high scoring detections. Thus, at level-2, the ego lane is wrongly predicted. Again, we are able to correct the error at level-3 (first image from right) when we jointly model the lanes.

Similarly, for the KITTI dataset depicted in Figure 4 (a) poses challenges due to complex shadow formations. Figures 4 (b,d,e) pose challenges due to the insufficient evidence from painted lane boundaries. We observe that despite noisy detections, level 3 has been able to achieve promising performance on detecting the lanes.

6.3. Scene Description

In contrast to existing approaches on lane/road detection [24, 3, 12, 25, 14] techniques, our approach is capable of providing an intuitive *scene description*. We underlay this assumption with the following examples: in Figure 5 (a) we observe (i) one truck on the ego lane; (ii) another truck on the left lane. In Figure 5 (b) (i) two pedestrians on the left side walk; (ii) one pedestrian on the right side walk. In Figure 5 (c), (i) one pedestrian is standing on the left side of the left-lane; (ii) one pedestrian is standing on the right side of the right cyclist lane; (iii) two cars on the left-lane; (iv) one car on the ego lane. In Figure 5 (d), (i) two cyclists on the right-lane; (ii) a car on the ego-lane. *Such intuitive scene descriptions can provide input to multiple driver assistance systems, including fully autonomous driving.*

7. Conclusion

Our proposed approach adopts a natural way to interpret a road scene in terms of a hierarchical part based model. We extend HPMs to scene understanding by proposing an

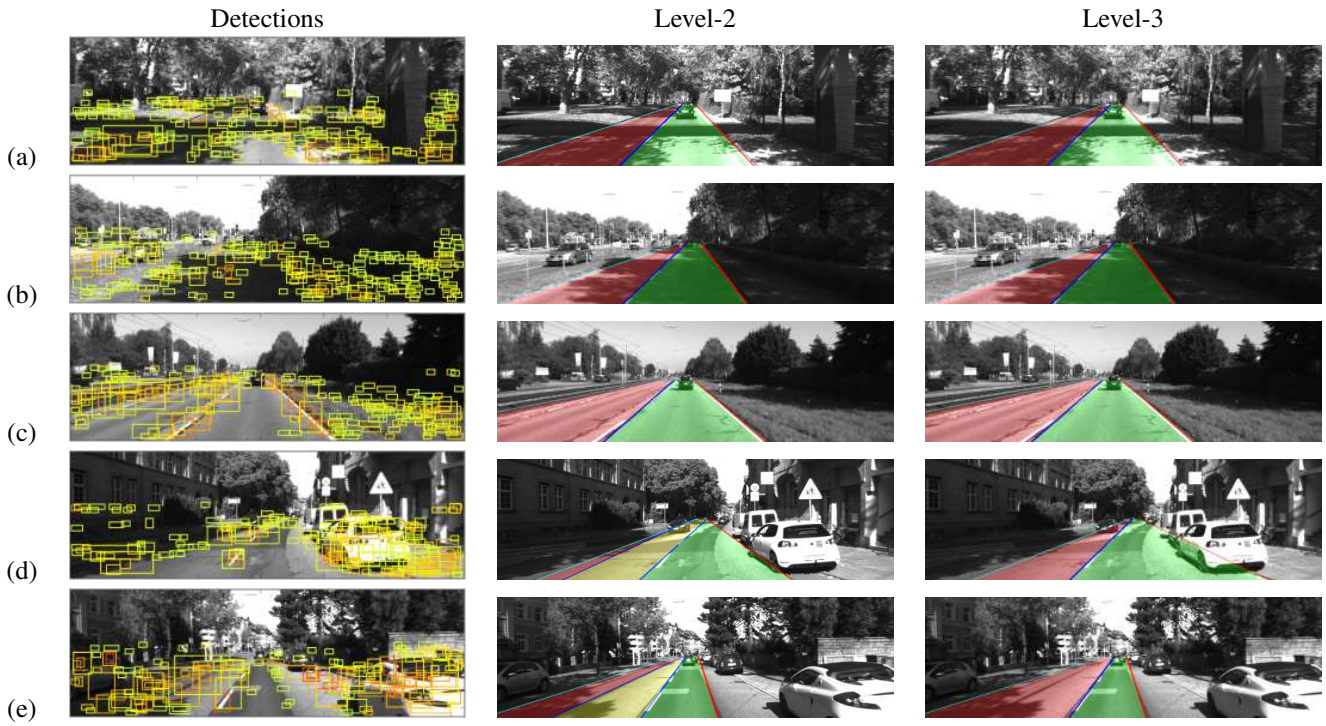


Figure 4. Qualitative results on images from KITTI dataset. Despite a large number of high scoring false positive detections at Level-0, Level-2 results in an optimal interpretation of the road region in terms of its constituent lanes in rows a,b and c. However in rows d and e, Level-2 has resulted in erroneous overlap (yellow) of the ego lane (green) and left lane (red) predictions. Level-3 further corrects the errors at Level-2 by taking spatial relationships between lanes into consideration.

approach that dispenses with the need for having root filters for non-leaf nodes, by treating them as *latent* variables within a DP-based optimization. Our experiments demonstrate that we can significantly reduce errors, when more information is included at multiple levels in a structured manner. Moreover, we find that the hierarchical interpretation also leads to intuitive scene descriptions that are central to autonomous driving.

In the future, we plan to investigate the combination of part based models with semantic segmentation based approach in the context of scene understanding, taking inspiration from recent work [29, 17, 13] along this direction. Furthermore, we also plan to extend the proposed LHPMs to model the dynamic aspects (*e.g.* tracking and interactions between scene entities) of a scene.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Discriminative appearance models for pictorial structures. *IJCV*. 1, 2, 3

[2] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, pages 44–57. 2008. 2

[3] G. Chen. *Texture Based Road Surface Detection*. PhD thesis, Case Western Reserve University, 2008. 6

[4] J. J. Corso. Toward parts-based scene understanding with pixel-support parts-sparse pictorial structures. *Pattern Recognition Letters*, 2013. 1

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 4

[6] A. Ess, T. Mueller, H. Grabner, and L. J. Van Gool. Segmentation-based urban traffic scene understanding. In *BMVC*, volume 1, 2009. 2

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 4

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 2010. 1, 2, 3

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005. 1, 2, 3

[10] J. Fritsch, T. Kuhnl, and A. Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *ITSC*, pages 1693–1700, 2013. 4, 6

[11] S. Gu, Y. Zheng, and C. Tomasi. Nested pictorial structures. In *ECCV*, pages 816–827, 2012. 2

[12] Y. He, H. Wang, and B. Zhang. Color-based road detection in urban traffic scenes. *ITSC*, 5(4):309–318, 2004. 6

[13] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, pages 30–43. 2008. 7

[14] H. Kong, J.-Y. Audibert, and J. Ponce. General road detection from a single image. *IEEE Trans. on Image Processing*, 19(8):2211–2220, 2010. 6

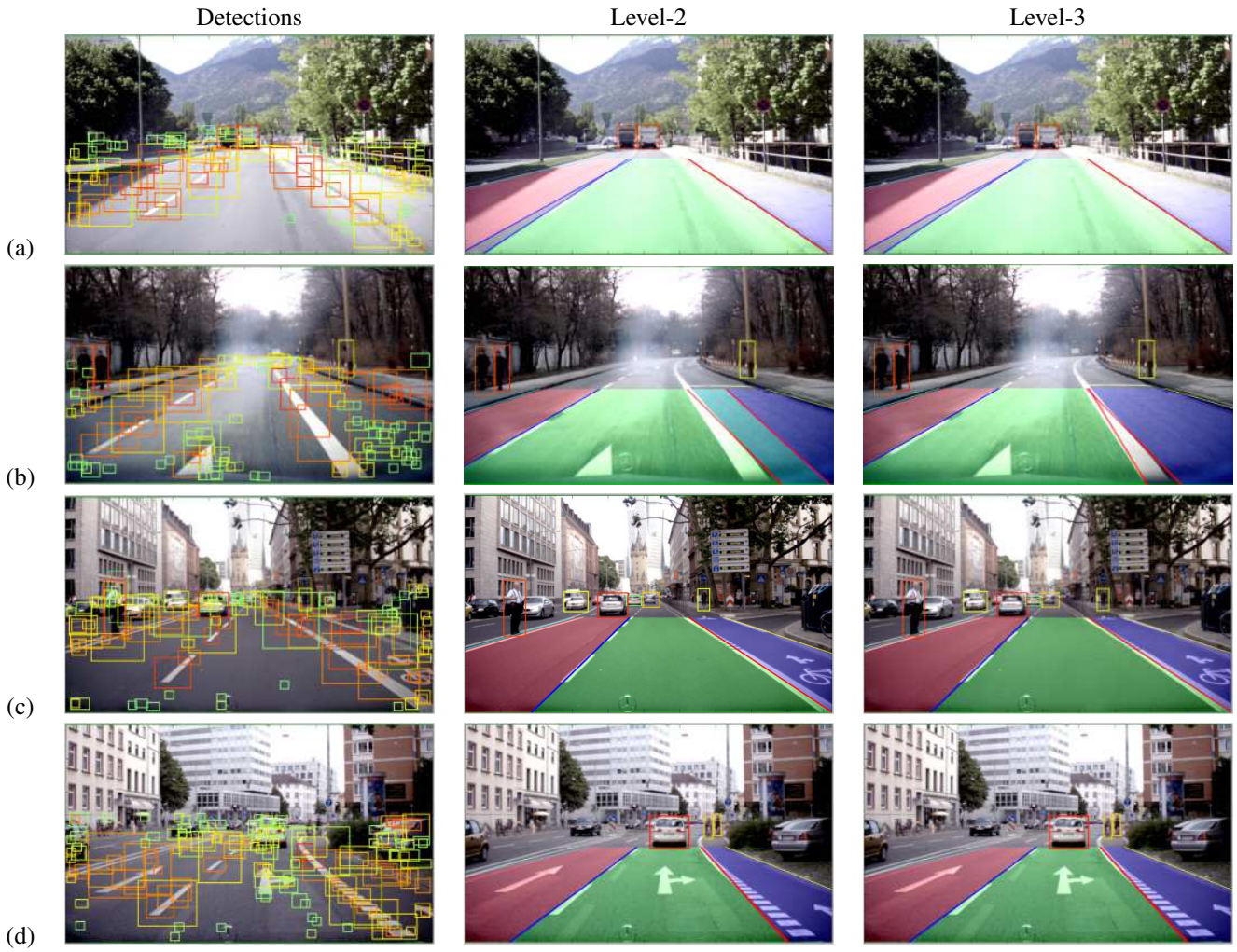


Figure 5. Qualitative results on images from Continental dataset. Despite a large number of high scoring false positive detections at Level-0, Level-2 results in an optimal interpretation of the road region in terms of its constituent lanes in rows a,c and d. However in row b Level-2 has resulted in an erroneous overlap (cyan) of the ego lane (green) and right lane (blue) predictions. Level-3 further corrects the errors at Level-2 by taking spatial relationships between lanes into consideration.

[15] T. Kühnl, F. Kummert, and J. Fritsch. Spatial ray features for real-time ego-lane extraction. In *ITSC*, 2012. 6

[16] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical random fields. *PAMI*, 36(6):1056–1077, 2014. 2

[17] L. Ladickỳ, P. Sturgess, K. Alahari, C. Russell, and P. H. Torr. What, where and how many? combining object detectors and crfs. In *ECCV*, pages 424–437. 2010. 7

[18] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. 2

[19] O. Miksik, D. Munoz, J. A. Bagnell, and M. Hebert. Efficient temporal consistency for streaming video scene analysis. In *ICRA*, pages 133–139, 2013. 2

[20] D. Munoz, J. A. Bagnell, and M. Hebert. Stacked hierarchical labeling. In *ECCV*, pages 57–70. 2010. 2

[21] P. Ott and M. Everingham. Shared parts for deformable part-based models. In *CVPR*, 2011. 1, 2, 3

[22] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, pages 1307–1314, 2011. 1, 2

[23] G. Roig, X. Boix, F. De la Torre, J. Serrat, and C. Vilella. Hierarchical crf with product label spaces for parts-based models. In *FG*, pages 657–664, 2011. 2

[24] A. G. Stevica Graovac. Detection of road image borders based on texture classification. *IJARS*, 2012. 6

[25] C. Tan, T. Hong, T. Chang, and M. Shneier. Color model-based real-time learning for road following. In *ITSC*, pages 939–944, 2006. 6

[26] D. Topfer, J. Spehr, J. Effertz, and C. Stiller. Efficient road scene understanding for intelligent vehicles using compositional hierarchical models. *ITSC*, 16(1):441–451, 2015. 2

[27] V. Vladimir and A. Lerner. Pattern recognition using generalized portrait method. 1963. 4

[28] C. Wojek and B. Schiele. A dynamic conditional random field model for joint labeling of object and scene classes. In *ECCV*, pages 733–747. 2008. 2

[29] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, pages 702–709, 2012. 7

[30] Y. Yi and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392, 2011. 1, 2, 3

[31] S. Zuffi, O. Freifeld, and M. J. Black. From pictorial structures to deformable structures. In *CVPR*, pages 3546–3553, 2012. 1, 2, 3