Gibbs Sampling for 2D Cane Structure Extraction From Images

Ricardo D. C. Marin
Department of Computer Science
University Of Canterbury
Canterbury, Christchurch
ricardo.castanedamarin@pg.canterbury.ac.nz

Tom Botterill
Department of Computer Science
University Of Canterbury
Canterbury, Christchurch

Richard D. Green
Department of Computer Science
University Of Canterbury
Canterbury, Christchurch

***Abstract*—In this paper we are interested in recovering 2D tree structure of vines from binary images. We propose a bottom-up approach that firstly segments an input image into cane parts, and second infer their connectivity by using Gibbs Sampling. Our approach is similar to previous work on vine structure inference [1], but instead of the use of heuristics for connecting cane parts, our method uses Gibbs sampling which has been successfully used in similar computer vision tasks [2]. We show comparative results against [1], and we provide directions on how this work could be extended in the future.**

I. Introduction

A vine pruning robot is using computer vision to reconstruct 3D model of vines and decide which canes should be cut [1]. The 3D models are found by matching 2D structures of vines that are extracted from images taken at different angles. Our system then decides which canes to leave by analysing the 3D models and builds a path that the robot arm follows to make the cuts. In this paper we focus on the extraction of 2D canes structure from grape vine images. Our goal is to recover the tree structure present in vine images. For this, all canes must be found, including their position and connections between them.

Recovering 2D structures of vines from images is a complex task. Figure 1 shows some of the vine image regions that make estimating its hierarchy a hard problem. Similar to Botterill et. al [1] our solution uses a bottom-up parsing of the vine image. However, instead of having as primitives edge segments, we use a custom cane segmentation based on a binary scanning algorithm [3] as show in Figure 2. Our primitives are then cane segments with constant width and defined local orientation. Finally, we use Gibbs sampling to infer connectivity.

The main contributions of this paper are firstly a detailed description of applying Gibbs sampling and cane segmentation to recover vine 2D structures; and secondly a quantitative comparison of our method with previous work on 2D cane structure extraction [1].

This paper is structured as follows. Section II reviews the background theory relevant to our method, in particular bottom-up/top-down frameworks are briefly reviewed. Section III presents our problem formulation and describes how to use Gibbs Sampling to infer the structure of the vines. In Section IV we show results of applying our method in comparison with the current method being used in the vine pruning system [1]. Finally, Section V presents a discussion and describe ideas that could be used for improving our method in the future.

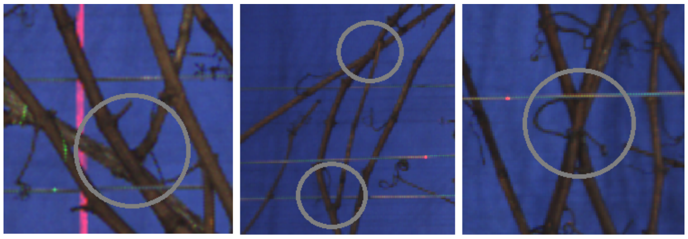

Fig. 1. Different cane occlusions and connectivity issues. From left to right: Branching occluded by another cane; cane tips finishing at non background regions; and multiple canes occluded in the same region.

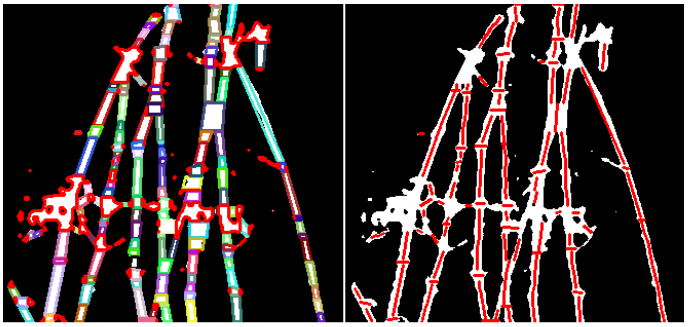

Fig. 2. Cane segmentation using a binary image scanning algorithm [3]. Left: Extracted cane segments; Right: Line between the two extreme points defined for each cane segment. See [3] for a detailed explanation.

II. Related Research

This research is related to bottom-up/top-down approaches and part-based modeling in computer vision [2], [4]–[12]. In bottom-up/top-down methods, an image is decomposed hierarchically in constituent visual patterns or primitives in a bottom up manner [2]. Then a model that encapsulate relations between these primitives is built, giving a semantic interpretation of an input image [6]. Similarly in part-based models, an articulated object is learn as relative positioning and other relations between its hierarchy parts [11].

Roughly speaking, we can decompose these frameworks into two fundamental components [2]. The first component has to do with obtaining the set of primitives or image parts that we want in our hierarchical model. These standard elements are based on edges or contour descriptors, segmented parts of the image, shape descriptors, or even whole detected individual objects [6]. Thus methods for this component usually relates to standard image processing and vision procedures like segmentation and recognition [2], [8]. For vine structure retrieval, Botteril et. al [1] used contours to build up a set of edge segments to represent vines canes. In contrast to this, our method is based on the cane segmentation described in [3]. This segmentation builds a set of vine clusters with properties that models real cane segments, like smoothly varying width and curvature, and consistent local orientation.

The second component has to do with modeling the relations between the set of primitives. Here most works can be further classified into two main lines of research. The first models individual articulated objects like humans, cars, bicycles etc. [5], [11], [13], [14]. The second models high level relationships between whole objects like for example a car in a street, or a human riding a bicycle [8], [10], [12], [15]. We note that Image Grammars [6] is a framework that is able to reason with individual objects representations, and relations between them at the same time.

Given that relations between the primitives or image parts in question are usually encoded as graphical models, in general hierarchical modeling is related to MAP – *Maximum a Posteriori* estimation in vision problems and inference in probabilistic models [5], [6]. More specifically, given a set of primitives or image parts of an input image, a configuration x that establish relations between them is usually found by maximizing the conditional distribution $P(x|I)$ of the configuration given an observed image I [2], [5]. In the Computer Vision literature, methods that have been used for MAP estimation are varied including sampling techniques like Monte Carlo, or optimization frameworks like simulated annealing, or gradient based methods [16]. Usually, the dimension of the configuration space is so high that brute force searches are not computationally feasible [2], and MAP inference is sometimes an NP hard problem [16].

Another way of treating the relationship between primitives or image parts is to use machine learning [1], [11]. In Felzenszwalb et. al [11], a deformable part model is learned by using a database of manually annotated bounding boxes around the object in question and by reducing the problem to binary classification that is solved with a modified version of Support Vector Machines (SVM). Similarly, for cane structure extraction, Botterill et. al [1] poses the problem of connecting two cane edge segments as a binary classification problem. The classifier used is SVM. Other classifiers such as random forest or neural networks could be trained and used as well [1].

Finally our method makes use of MAP estimation rather than machine learning. In particular, we model connections between cane segments as binary variables. Then we use Gibbs sampling to sample from a joint distribution of all the binary variables, effectively recovering the most likely connectivity for our cane segments given the image data. The method is described in the following sections.

III. Cane Structure Extraction From An Image

Our cane structure extraction system is divided into two main subsystems, cane segmentation and structure inference. Both subsystems are described in the following.

A. Cane Segmentation

In this subsystem we are given a binary image I and our objective is to find a segmentation $C = \{s_k\}$, $k = 1, ..., n$ of the vine pixels, such that each s_k have well defined local orientation, and smoothly changing width and curvature. These properties have been chosen in order for the cane segments s_k to match real cane parts. Figure 2 shows an example of a cane segmentation obtained by applying the binary image scanning algorithm [3]. Each cane segment has an associated 2D curve shown red in this figure. We will denote by p_k^e with $e = 1, 2$ the two extreme points of this curve for each segment s_k. These points act like "bonds" defined in image grammars to connect visual words [6]. For further details on how to obtain a cane segmentation with the desired properties see [3]. The structure inference system described in the next section is independent of this cane segmentation, and it can be used with other segments, with the mentioned properties.

B. Structure Inference

In this subsystem, we have a cane segmentation C as described in Section III-A, and we are interested in reconstructing from it the tree structure inherent in the imaged vine. Observe that this tree like structure can be modeled as a set of connections between the extreme points p_k^e of the cane segments. In this case, vine structure extraction can be done by selecting among the set of all possible connections the one that satisfy special properties related to vine images, for example, connection of cane segments of similar thickness and smooth angle variations between them.

Formally, given a pair of extreme points p_k^e and $p_{k'}^{e'}$ of cane segments $s_k, s_{k'}$, we model connectivity between the extreme points by defining a binary variable

$$x((k,e),(k',e')) = \begin{cases} 1 & p_k^e \text{ is connected to } p_{k'}^{e'} \\ 0 & \text{otherwise} \end{cases}$$

Now denote $\mathbf{x} = (x_1, x_2, ..., x_m)$ the concatenation of all binary variables of all extreme points of all cane segments. We will use the index $i = 1, ..., m$ to enumerate all of the components of $\mathbf{x}$. Observe that an instance of $\mathbf{x}$ encodes a configuration of all cane segments being connected and disconnected at their extreme points. Thus for vine structure inference we are interested in finding a $\mathbf{x}^*$ such that

$$\mathbf{x}^* = \max_{\mathbf{x}} P(\mathbf{x}|I) \qquad (1)$$

where P is the probability distribution of $\mathbf{x}$ given the observed vine image I, which should be modeled to carry vine's special properties. This reduces our problem of finding cane structure to that of MAP estimation, and we can make use of sampling techniques [17] to estimate $\mathbf{x}^*$. Gibbs sampling [17] is a method that enable us to sample from an unknown joint distribution $P(x_1, ..., x_m|I)$ by using samples of the conditional distributions $P(x_i|x_1, ...x_{i-1}, x_{i+1}, ..., x_m, I) := P(x_i|\mathbf{x}_{-i}, I)$ and the sequence of samples under regular conditions would become samples of [18], [19]. Algorithm 1 summarize this sampling procedure. We initialize our model to have all pairs of candidates disconnected. Then, we compute iteratively the probability of a single connection x_i given we

```
Data: Binary image $I$, cane segmentation $C$,
      maximum number of iterations T.
Result: Estimate of $\mathbf{x}^*$
Initialize $\mathbf{x}^{(0)}$ to string of zeros;
for $t \leftarrow 1$ to T do
    for $i \leftarrow 1$ to $m$ do
        compute $p = P(x_i^{(t)} | \mathbf{x}_{-i}^{(t-1)}, I)$;
        sample $u = \mathrm{Unif}(0, 1)$;
        if $u < p$ then
            $x_i^{(t)} = 1$
        else
            $x_i^{(t)} = 0$
        end
    end
end
set $\mathbf{x}^* = x_T$;
```

Algorithm 1: Gibbs Sampling.

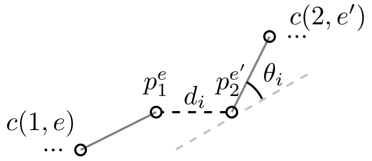

Fig. 3. Attributes for a connection between a pair of points p_1^e and $p_2^{e'}$. For each point we can use their associated canes $c(1, e)$ and $c(2, e')$ to compute the angle θ_i between the two segments. We also use the Euclidean distance d_i between both points.

	G. Truth	Found	Correct	Precision	Recall
		80% Canes Overlap			
Gibbs	1093	10348	1009	0.0975068	0.923147
Contour	1093	1355	746	0.550554	0.682525
		50% Canes Overlap			
Gibbs	1093	10348	1032	0.0997294	0.94419
Contour	1093	1355	859	0.633948	0.78591
		50% Canes Length Coverage			
Gibbs	1093	10348	1.47e+05	0.141376	0.303813
Contour	1093	1355	1.90e+05	0.510073	0.390673

TABLE I. Cane Structure Matching To Ground Truth

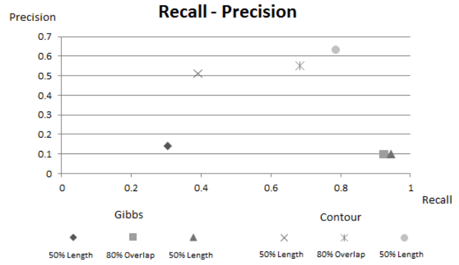

Fig. 4. Comparative results of extracting canes using our method and the contour method in Botterill et. al [1].

know all other cane connectivities $\mathbf{x}_{-i}$. We then update the connection x_i as a sample of a Bernoulli distribution with parameter p. We repeat this process for all i and we are guaranteed, under regular conditions, that when the number of iterations $\mathrm{T} \to \infty$ the samples $\mathbf{x}^{(t)}$ will belong to the model posterior $P(\mathbf{x}|I)$ [17]. The first samples $\mathbf{x}^{(t)}$ may be biased, so it is common to choose a value $t = b$, and estimate $\mathbf{x}^*$ using only $t \geq b$. This is called the burning-in period in the Gibbs sampling framework [18]. Therefore, all that rest here is to define $P(x_i|\mathbf{x}_{-i}, I)$ such that it satisfy our vine data.

C. Modeling $P(x_i|\mathbf{x}_{-i}, I)$

Given a point p_k^e we restrict the set of candidate points that can potentially be connected to this point, by filtering to the points that are within a radius R. We then define a set of attributes that allow us to model good connections relative to vine's structure. Suppose x_i represents the connection between points p_k^e and $p_{k'}^{e'}$. Given that we know all connectivity states $\mathbf{x}_{-i}$, we are able to reconstruct canes up to the connection represented by x_i. Denote by $c(k, e)$ the cane passing through the point p_k^e. For each cane $c(k, e)$, we are able to extract a set of attributes that we use for modeling $P(x_i|\mathbf{x}_{-i}, I)$. In particular we use the angle θ_i between the segments of each $c(k, e)$ that ends in p_k^e or $p_{k'}^{e'}$, as shown in Figure 3. We also use the Euclidean distance d_i between these points. Similar to Botterill et. al [19] we model angles between positive connections by a normal probability density $p(\theta_i) = \mathcal{N}(\mu, \sigma)$ whose parameters can be learned from ground truth data. Distances d_i of positive connections are modeled with a uniform distribution $p(d_i)$. Analogously, we can model the distribution of these attributes but now for negative connections $p'(\theta_i)$ and $p'(d_i)$. We used uniform distributions in our experimental results. Now denote by $a_i = (\theta_i, d_i)$ the concatenation of the attributes and $p(a_i) := p(\theta_i, d_i) = p(\theta_i)p(d_i)$, $p'(a_i) := p'(\theta_i, d_i) = p'(\theta_i)p'(d_i)$. Then we can compute $P(x_i|\mathbf{x}_{-i}, I) = P(x_i|a_i)$ by using Bayes formula [19]:

$$P(x_i|a_i) = \frac{P(x_i)p(a_i)}{P(x_i)p(a_i) + (1 - P(x_i))p'(a_i)}$$

where $P(x_i)$ represents the prior of the connection x_i. Observe that this estimation could be easily extended to use more attributes between canes c_k.

IV. Experimental Results

We assessed our method of Gibbs sampling experimentally on real vine images that have been manually annotated with ground truth canes. Figure 5 shows results of canes extracted using our Gibbs sampling method. We measured precision and recall for extracted canes matching to this ground truth. A cane is considered correct if it overlaps with $p\%$ of a ground truth cane, or if it covers $p\%$ of the length of the ground truth cane. We used $p = 80\%$ and $p = 50\%$ for overlaps and $p = 50\%$ for length coverage. To evaluate our precision and recall results, we compared our approach to the contour method by Botterill et. al [1]. All results are shown in Table I and Figure 4. We have found that our method suffers from low precision due to the high amount of single canes that do not get connected together. On the other hand, we achieved higher recall than the contour method, with 89% and 91% of our canes matching correctly the ground truth data. We also noted a low precision in the length coverage. This is a consequence again that though the canes overlaps satisfactorily to the ground truth

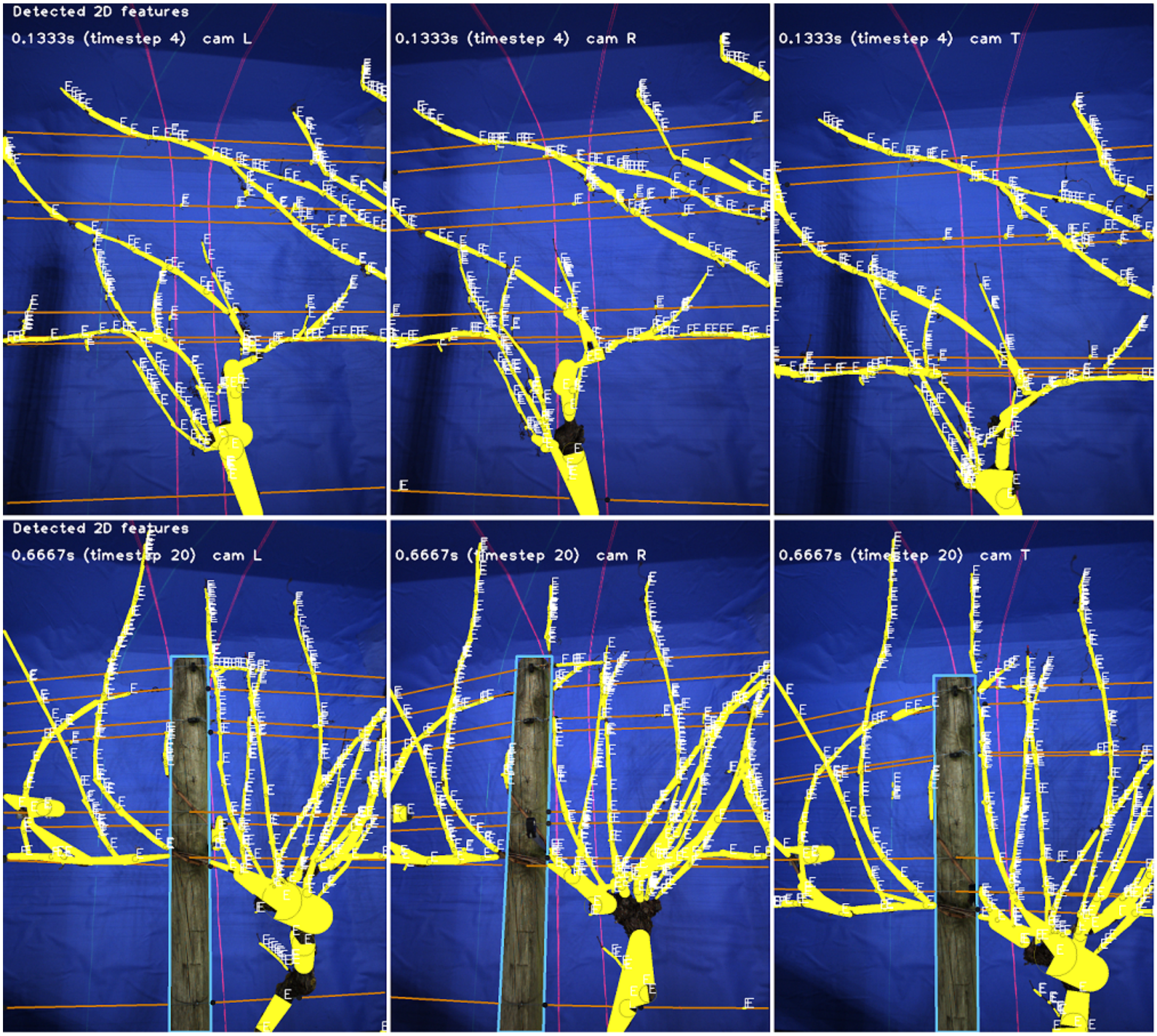
Detected 2D features
0.1333s (timestep 4) cam L
0.1333s (timestep 4) cam R
0.1333s (timestep 4) cam T
Detected 2D features
0.6667s (timestep 20) cam L
0.6667s (timestep 20) cam R
0.6667s (timestep 20) cam T

Fig. 5. Canes found by using Gibbs sampling or real vine images. From left to right, on each frame the pruning system extracts the vine structure from three different perspectives of the same vine. From top to bottom, to different frames of the system. An *E* in the image represents the ending point of a cane. As we can see many of the canes remain unconnected.

data, canes recovered by our Gibbs sampling method are only approximating truth canes by parts.

V. Conclusion and Future Work

This paper presented an application of Gibbs sampling to recover canes from vine images. The approach was based on a cane segmentation of a vine binary map and connectivity inference by sampling from a posterior distribution that modeled the connections between different cane segments. We applied our algorithm to real vine data and compare our method to the one currently being used in our system [1]. We have found our method improves recall but suffers from low precision compared to the contour method [1].

Limitations of our approach are the inability to detect branching points. This is because the connectivity model does not allow a node point to be connected to multiple other node points. In the future we aim to classify each point as either a cane tip, a branch point or a connection point. The connectivity model would take into consideration this information and may incorporate another model for relations between different types of points similar to image grammars [6]. Another limitation of our approach is that is not able to detect cycles in the connectivity graph. In the future we aim to address this by taking into account these cycles on the computation of the conditional probabilities of Section III-C.

Another way to improve our approach would be to use more attributes for interactions between cane segments. In particular canes overlapping connectivity could be enhanced by using color information rather than only vine pixels of the binary image. Contour extraction in these regions and connectivity as modeled in Botterill et. al [1] could be used together with Gibbs sampling. Finally, alternative methods for MAP estimation and other structured models could be researched for vine structure extraction.

REFERENCES

[1] T. Botterill, R. Green, and S. Mills, “Finding a vine’s structure by bottom-up parsing of cane edges,” in *Image and Vision Computing New Zealand (IVCNZ), 2013 28th International Conference of*, Nov 2013, pp. 112–117.

[2] Z. Tu, X. Chen, A. Yuille, and S.-C. Zhu, “Image parsing: Unifying segmentation, detection, and recognition,” *International Journal of Computer Vision*, vol. 63, no. 2, pp. 113–140, 2005. [Online]. Available: http://dx.doi.org/10.1007/s11263-005-6642-x

[3] R. Marin, T. Botterill, and R. Green, “Binary image scanning algorithm for cane segmentation,” University of Canterbury, Tech. Rep., 2014. [Online]. Available: www.hilandtom.com/ricardo.pdf

[4] T. Matsuyama and V. Hwang, “Sigma: A framework for image understanding integration of bottom-up and top-down analyses,” in *Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI’85. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1985, pp. 908–915. [Online]. Available: http://dl.acm.org/citation.cfm?id=1623611.1623659

[5] P. F. Felzenszwalb and D. P. Huttenlocher, “Pictorial structures for object recognition,” *Int. J. Comput. Vision*, vol. 61, no. 1, pp. 55–79, Jan. 2005. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000042934.15159.49

[6] S.-C. Zhu and D. Mumford, “A stochastic grammar of images,” *Found. Trends. Comput. Graph. Vis.*, vol. 2, no. 4, pp. 259–362, Jan. 2006. [Online]. Available: http://dx.doi.org/10.1561/0600000018

[7] F. Han and S.-C. Zhu, “Bottom-up/top-down image parsing with attribute grammar,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 59–73, Jan 2009.

[8] Y. Zhao and S.-C. Zhu, “Image parsing via stochastic scene grammar,” *NIPS, Advances in Neural Information Processing Systems*, 2011.

[9] R. B. Girshick, P. F. Felzenszwalb, and D. Mcallester, “Object detection with grammar models,” in *In NIPS*, 2011.

[10] J. Tighe and S. Lazebnik, “Finding things: Image parsing with regions and per-exemplar detectors,” in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR ’13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 3001–3008. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2013.386

[11] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[12] J. Malik, L. Bourdev, S. Gupta, C. Gu, B. Hariharan, and P. Arbelaez, “Semantic segmentation using regions and parts,” *2013 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 3378–3385, 2012.

[13] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, Sept 2010.

[14] P. F. Felzenszwalb, R. B. Girshick, and D. Mcallester, “D.m.: Cascade object detection with deformable part models,” in *In: Proc. CVPR. (2010*.

[15] G. Floros, K. Rematas, and B. Leibe, “B.: Multi-class image labeling with top-down segmentation and generalized robust p?n potentials,” in *In: BMVC. (2011*.

[16] C. Wang and N. Paragios, “Markov random fields in vision perception: A survey,” INRIA, Tech. Rep., 2012.

[17] G. Heitz, “Graphical models for high-level computer vision,” Ph.D. dissertation, Stanford, CA, USA, 2009.

[18] *Markov Chain Monte Carlo in Practice (Chapman & Hall/CRC Interdisciplinary Statistics)*, Softcover reprint of the original 1st ed. 1996 ed. Chapman and Hall/CRC, Dec. 1995.

[19] T. Botterill, R. Green, and S. Mills, “A decision-theoretic formulation for sparse stereo correspondence problems,” in *To appear in Proc. 3DV*, 2014.