Parsimonious Labeling

Puneet K. Dokania
CentraleSupélec and INRIA Saclay
puneet.kumar@inria.fr

M. Pawan Kumar
CentraleSupélec and INRIA Saclay
pawan.kumar@ecp.fr

Abstract

We propose a new family of discrete energy minimization problems, which we call parsimonious labeling. Our energy function consists of unary potentials and high-order clique potentials. While the unary potentials are arbitrary, the clique potentials are proportional to the diversity *of the set of unique labels assigned to the clique. Intuitively, our energy function encourages the labeling to be parsimonious, that is, use as few labels as possible. This in turn allows us to capture useful cues for important computer vision applications such as stereo correspondence and image denoising. Furthermore, we propose an efficient graph-cuts based algorithm for the parsimonious labeling problem that provides strong theoretical guarantees on the quality of the solution. Our algorithm consists of three steps. First, we approximate a given diversity using a mixture of a novel hierarchical P^n Potts model. Second, we use a divide-and-conquer approach for each mixture component, where each subproblem is solved using an efficient α-expansion algorithm. This provides us with a small number of putative labelings, one for each mixture component. Third, we choose the best putative labeling in terms of the energy value. Using both synthetic and standard real datasets, we show that our algorithm significantly outperforms other graph-cuts based approaches.*

1. Introduction

The labeling problem provides an intuitive formulation for several tasks in computer vision and related areas. Briefly, the labeling problem is defined using a set of random variables, each of which can take a value from a finite and discrete label set. The assignment of values to all the variables is referred to as a labeling. In order to quantitatively distinguish between the large number of putative labelings, we are provided with an energy function that maps a labeling to a real number. The energy function consists of two types of terms: (i) the unary potential, which depends on the label assigned to one random variable; and (ii) the clique potential, which depends on the labels assigned to a subset of random variables. The goal of the labeling problem is to obtain the labeling that minimizes the energy.

A well-studied special case of the labeling problem is metric labeling [2, 16]. Here, the unary potentials are arbitrary. However, the clique potentials are specified by a user-defined metric distance function of the label space. Specifically, the clique potentials satisfy the following two properties: (i) each clique potential depends on two random variables; and (ii) the value of the clique potential (also referred to as the pairwise potential) is proportional to the metric distance between the labels assigned to the two random variables. Metric labeling has been used to formulate several problems in low-level computer vision, where the random variables correspond to image pixels. In such scenarios, it is natural to encourage two random variables that correspond to two nearby pixels in the image to take similar labels. However, by restricting the size of the cliques to two, metric labeling fails to capture more informative high-order cues. For example, it cannot encourage an arbitrary sized set of similar pixels (such as pixels that define a homogeneous superpixel) to take similar labels.

We propose a natural generalization of the metric labeling problem for high-order potentials, which we call parsimonious labeling. Similar to metric labeling, our energy function consists of arbitrary unary potentials. However, the clique potentials can be defined on any subset of random variables, and their value depends on the set of unique labels assigned to the random variables in the clique. In more detail, the clique potential is defined using the recently proposed notion of a *diversity* [3, 4], which generalizes metric distance functions to all subsets of the label set. By minimizing the diversity, our energy function encourages the labeling to be parsimonious, that is, use as few labels as possible. This in turn allows us to capture useful cues for important low-level computer vision applications.

In order to be practically useful, we require a computationally feasible solution for parsimonious labeling. To this end, we design a novel three step algorithm that uses an efficient graph cuts based method as its key ingredient. The first step of our algorithm approximates a given diversity as a mixture of a novel *hierarchical P^n Potts model* (a generalization of the P^n Potts model [17]). The second step of our algorithm solves the labeling problem correspond-

ing to each component of the mixture via a divide-and-conquer approach, where each subproblem is solved using α-expansion [29]. This provides us with a small set of putative labelings, each corresponding to a mixture component. The third step of our algorithm chooses the putative labeling with the minimum energy. Using both synthetic and real datasets, we show that our overall approach provides accurate results for various computer vision applications.

2. Related Work

In last few years the research community has witnessed many successful applications of high-order random fields to solve low level vision related problems such as object segmentation [8, 9, 18, 22, 28, 30], disparity estimation [15, 31], and image restoration [23]. In this work, our focus is on methods that (i) rely on efficient move-making algorithms based on graph cuts; (ii) provide a theoretical guarantee on the quality of the solution. Below, we discuss the work most closely related to ours in more detail.

Kohli et al. [17] proposed the P^n Potts model, which enforces label consistency over a set of random variables. In [18], they presented a robust version of the P^n Potts model that takes into account the number of random variables that have been assigned an inconsistent label. Both the P^n Potts model and its robust version lend themselves to the efficient $\alpha-$expansion algorithm [17, 18]. Furthermore, the $\alpha-$expansion algorithm also provides a multiplicative bound on the energy of the estimated labeling with respect to the optimal labeling. While the robust P^n Potts model has been shown to be very useful for semantic segmentation, our generalization of the P^n Potts model offers a natural extension of the metric labeling problem and is therefore more widely applicable to several low-level vision tasks.

Delong et al. [8] proposed a global clique potential (label cost) that is based on the cost of using a label or a subset of labels in the labeling of the random variables. Similar to the P^n Potts model, the label cost based potential can also be minimized using $\alpha-$expansion. However, the theoretical guarantee provided by $\alpha-$expansion is an additive bound, which is not invariant to reparameterization of the energy function. Delong et al. [7] also proposed an extension of their work to hierarchical costs. However, the assumption of a given hierarchy over the label set limits its applications.

Ladicky *et al.* [22] proposed a global co-occurrence cost based high order model for a much wider class of energies that encourage the use of a small set of labels in the estimated labeling. Theoretically, the only constraint that [22] enforces in high order clique potential is that it should be monotonic in the label set. In other words, [22] can be regarded as a generalization of parsimonious labeling. However, they approximately optimize an upperbound on the actual energy function, which does not provide any optimality guarantees. In our experiments, we demonstrate that our move-making algorithm significantly outperforms their approach for the special case of parsimonious labeling.

Fix *et al.* [12] proposed an algorithm (SoSPD) for high-order random fields with arbitrary clique potentials. Each move of this algorithm requires us to approximately upperbound the clique potential into a submodular function and then optimize it using the submodular max-flow algorithm [19]. We show that our move making algorithm for parsimonious labeling has a much stronger multiplicative bound and better time complexity compared to [12].

3. Preliminaries

The labeling problem. Consider a random field defined over a set of random variables $\mathbf{x} = \{x_1, \cdots, x_N\}$ arranged in a predefined lattice $\mathcal{V} = \{1, \cdots, N\}$. Each random variable can take a value from a discrete label set $\mathcal{L} = \{l_1, \cdots, l_H\}$. Furthermore, let $\mathcal{C}$ denote the set of maximal cliques. Each maximal clique consists of a set of random variables that are all connected to each other in the lattice. A labeling is defined as the assignment or mapping of random variables to the labels. To assess the quality of each labeling $\mathbf{x}$ we define an energy function as:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \theta_i(x_i) + \sum_{c \in \mathcal{C}} \theta_c(\mathbf{x}_c). \tag{1}$$

where $\theta_i(x_i)$ is the unary potential of assigning a label x_i to the variable i, and $\theta_c(\mathbf{x}_c)$ is the clique potential for assigning the labels $\mathbf{x}_c$ to the variables in the clique c. Clique potentials are assumed to be non-negative. As will be seen shortly, this assumption is satisfied by the new family of energy functions proposed in our paper. The total number of putative labelings is H^N, each of which can be assessed using its corresponding energy value. Within this setting, the labeling problem is to find the labeling corresponding to the minimum energy according to the function (1). Formally, the labeling problem is: $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} E(\mathbf{x})$.

P^n Potts model. An important special case of the labeling problem, which will be used throughout this paper, is defined by the P^n Potts model [17]. The P^n Potts model is a generalization of of the well known Potts model [24] for high-order energy functions (cliques can be of arbitrary sizes). For a given clique, the P^n Potts model is defined as:

$$\theta_c(\mathbf{x}_c) \propto \begin{cases} \gamma^k, & \text{if } x_i = l_k, \forall i \in c, \\ \gamma^{max}, & \text{otherwise,} \end{cases} \tag{2}$$

where γ^k is the cost of assigning all the nodes to the label $l_k \in \mathcal{L}$, and $\gamma^{max} > \gamma_k, \forall l_k \in \mathcal{L}$. Intuitively, the P^n Potts model enforces label consistency by assigning the cost of γ^{max} if there are more than one label in the given clique.

α-expansion for P^n Potts model. In order to solve the labeling problem corresponding to the P^n Potts model, Kohli *et al.* [17] proposed to use the $\alpha-$expansion algorithm [29]. The $\alpha-$expansion algorithm starts with an initial labeling,

for example, by assigning each random variable to the label l_1. At each iteration, the algorithm moves to a new labeling by searching over a large *move space*. Here, the move space is defined as the set of labelings where each random variable is either assigned its current label or the label α. The key result that makes $\alpha-$expansion a computationally feasible algorithm for the P^n Potts model is that the minimum energy labeling within a move-space can be obtained using a single minimum st-cut operation on a graph that consists of a small number (linear in the size of the variables and the cliques) of vertices and arcs. The algorithm terminates when the energy cannot be reduced further for any choice of the label α. We refer the reader to [17] for further details.

Multiplicative Bound. An intuitive and commonly used measure of the accuracy of an approximation algorithm is the multiplicative bound. Formally, the multiplicative bound of a given algorithm is said to be B if the following condition is satisfied for all possible values of unary potential $\theta_i(.)$, and clique potentials $\theta_c(\mathbf{x}_c)$:

$$\sum_{i\in\mathcal{V}}\theta_i(\hat{x}_i)+\sum_{c\in\mathcal{C}}\theta_c(\hat{\mathbf{x}}_c)\leq\sum_{i\in\mathcal{V}}\theta_i(x_i^*)+B\sum_{c\in\mathcal{C}}\theta_c(\mathbf{x}_c^*). \quad (3)$$

Here, $\hat{\mathbf{x}}$ is the labeling estimated by the algorithm and $\mathbf{x}^*$ is a globally optimal labeling. By definition of an optimal labeling (one that has the minimum energy), the multiplicative bound will always be greater than or equal to one [20]. The $\alpha-$expansion algorithm for the P^n Potts model has the multiplicative bound of $\lambda\min(\mathcal{M},|\mathcal{L}|)$, where, $\mathcal{M}$ is the size of the largest clique, $|\mathcal{L}|$ is the number of labels, and $\lambda=\frac{\gamma^{max}}{\gamma^{min}}$, where $\gamma^{min}=\min_{\mathbf{x}_c,\theta_c(\mathbf{x}_c)\neq 0}\theta_c(\mathbf{x}_c)$ and $\gamma^{max}=\max_{\mathbf{x}_c}\theta_c(\mathbf{x}_c)$[14].

4. Parsimonious Labeling

The parsimonious labeling problem is defined using an energy function that consists of unary potentials and clique potentials defined over cliques of arbitrary sizes. While the parsimonious labeling problem places no restrictions on the unary potentials, the clique potentials are specified using a *diversity* function [3]. Before describing the parsimonious labeling problem in detail, we briefly define the diversity function for the sake of completion.

Definition 1. *A diversity is a pair $(\mathcal{L},\delta)$, where $\mathcal{L}$ is the label set and δ is a non-negative function defined on subsets of $\mathcal{L}$, $\delta:\Gamma\rightarrow\mathbb{R},\forall\Gamma\subseteq\mathcal{L}$, satisfying :*

- *Non Negativity: $\delta(\Gamma)\geq 0$, and $\delta(\Gamma)=0$, if and only if, $|\Gamma|\leq 1$.*
- *Triangular Inequality: if $\Gamma_2\neq\emptyset$, $\delta(\Gamma_1\cup\Gamma_2)+\delta(\Gamma_2\cup\Gamma_3)\geq\delta(\Gamma_1\cup\Gamma_3),\forall\Gamma_1,\Gamma_2,\Gamma_3\subseteq\mathcal{L}$.*
- *Monotonicity: $\Gamma_1\subseteq\Gamma_2$ implies $\delta(\Gamma_1)\leq\delta(\Gamma_2)$.*

Using a diversity function, we can define a clique potential as follows. We denote by $\Gamma(\mathbf{x}_c)$ the set of unique labels in the labeling of the clique c. Then, $\theta_c(\mathbf{x}_c)=w_c\delta(\Gamma(\mathbf{x}_c))$, where δ is a diversity function and w_c is the non-negative weight corresponding to the clique c. Formally, the parsimonious labeling problem amounts to minimizing the following energy function:

$$E(\mathbf{x})=\sum_{i\in\mathcal{V}}\theta_i(x_i)+\sum_{c\in\mathcal{C}}w_c\delta(\Gamma(\mathbf{x}_c)). \quad (4)$$

Therefore, given a clique $\mathbf{x}_c$ and the set of unique labels $\Gamma(\mathbf{x}_c)$ assigned to the random variables in the clique, the clique potential function for the parsimonious labeling problem is defined using $\delta(\Gamma(\mathbf{x}_c))$, where $\delta:\Gamma(\mathbf{x}_c)\rightarrow\mathbb{R}$ is a diversity function.

Intuitively, diversities enforces parsimony by choosing a solution with fewer unique labels from a set of equally likely solutions. This is an essential property in many vision problems, for example, in the case of image segmentation, we would like to see label consistency within superpixels in order to preserve discontinuity. Unlike the P^n Potts model the diversity does not enforce the label consistency very rigidly. It gives monotonic rise to the cost based on the number of labels assigned to the given clique.

An important special case of the parsimonious labeling problem is the *metric labeling problem*, which has been extensively studied in computer vision [2, 21] and theoretical computer science [5, 16]. In metric labeling, the maximal cliques are of size two (pairwise) and the clique potential function is a metric distance function defined over the labels. Recall that a distance function $d:\mathcal{L}\times\mathcal{L}\rightarrow\mathbb{R}$ is a metric if and only if: (i) $d(.,.)\geq 0$; (ii) $d(i,j)+d(j,k)\geq d(i,k),\forall i,j,k$; and (iii) $d(i,j)=0$ if and only if $i=j$. Notice that there is a direct link between the metric distance function and the diversities. Specifically, metric distance functions are diversities defined on subsets of size 2. In other words, diversities are the generalization of the metric distance function and boil down to a metric distance function if the input set is restricted to the subsets with cardinality of at most two. Another way of understanding the connection between metrics and diversities is that *every diversity induces a metric*. In other words, consider $d(l_i,l_i)=\delta(l_i)=0$ and $d(l_i,l_j)=\delta(\{l_i,l_j\})$. Using the properties of diversities, it can be shown that $d(\cdot,\cdot)$ is a metric distance function. Hence, in case of energy function defined over pairwise cliques, the parsimonious labeling problem reduces to the metric labeling problem.

In the remaining part of this section we talk about a specific type of diversity called *the diameter diversity*. We show its relation with the well known P^n Potts model. Furthermore, we propose a *hierarchical P^n Potts model* based on the diameter diversity defined over a hierarchical clustering (defined shortly). However, note that our approach is applicable to any general parsimonious labeling problem.

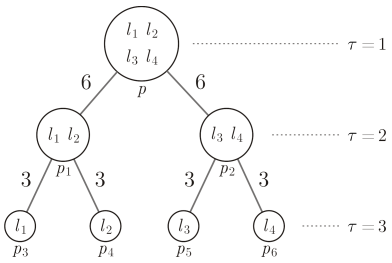


Figure 1: *An example of r-*HST *for $r = 2$. The cluster associated with root p contains all the labels. As we go down, the cluster splits into subclusters and finally we get the singletons, the leaf nodes (labels). The root is at depth of 1 ($\tau = 1$) and leaf nodes at $\tau = 3$. The metric defined over the r-*HST *is denoted as $d^t(.,.)$, the shortest path between the inputs. For example, $d^t(l_1, l_3) = 18$ and $d^t(l_1, l_2) = 6$. The diameter diversity for the subset of labels at cluster p is $\max_{\{l_i,l_j\}\in\{l_1,l_2,l_3,l_4\}} d^t(l_i, l_j) = 18$.*

Diameter Diversity. In this work, we are primarily interested in the diameter diversity [3]. Let $(\mathcal{L}, \delta)$ be a diversity and $(\mathcal{L}, d)$ be the induced metric of $(\mathcal{L}, \delta)$, where $d : \mathcal{L} \times \mathcal{L} \rightarrow \mathbb{R}$ and $d(l_i, l_j) = \delta(\{l_i, l_j\}), \forall l_i, l_j \in \mathcal{L}$, then for all $\Gamma \subseteq \mathcal{L}$, the diameter diversity is defined as:

$$\delta^{dia}(\Gamma) = \max_{l_i,l_j\in\Gamma} d(l_i, l_j). \tag{5}$$

Clearly, given the induced metric function defined over a set of labels, diameter diversity over any subset of labels gives the measure of the dissimilarity (or diversity) of the labels. More the dissimilarity, based on the induced metric function, higher is the diameter diversity. Therefore, using diameter diversity as clique potentials enforces the similar labels to be together. Thus, a special case of parsimonious labeling in which the clique potentials are of the form of diameter diversity can be defined as below:

$$E(\mathbf{x}) = \sum_{i\in\mathcal{V}} \theta_i(x_i) + \sum_{c\in\mathcal{C}} w_c \delta^{dia}(\Gamma(\mathbf{x}_c)). \tag{6}$$

Notice that the diameter diversity defined over uniform metric is nothing but the P^n Potts model where $\gamma_i = 0$. In what follows we define a generalization of the P^n Potts model, the hierarchical P^n Potts model, which will play a key role in the rest of the paper.

The Hierarchical P^n Potts Model. The hierarchical P^n Potts model is a diameter diversity defined over a special type of metric known as the r-HST metric. A rooted tree, as shown in Figure 1, is said to be an r-HST, or *r-hierarchically well separated* [1] if it satisfy the following properties: (i) all the leaf nodes are the labels; (ii) all edge weights are positive; (iii) the edge lengths from any node to all of its children are the same; and (iv) on any root to leaf path the edge weight decrease by a factor of at least $r > 1$. We can think of a r-HST as a hierarchical clustering of the given label set $\mathcal{L}$. The root node is the cluster at the top level of the hierarchy and contains all the labels. As we go down in the hierarchy, the clusters break down into smaller clusters until we get as many leaf nodes as the number of labels in the given label set. The metric distance function defined on this tree $d^t(.,.)$ is known as the r-HST metric. In other words, the distance $d^t(\cdot,\cdot)$ between any two nodes in the given r-HST is the length of the unique path between these nodes in the tree. *The diameter diversity defined over $d^t(.,.)$ is called the hierarchical P^n Potts model*. Figure 1 shows an example of diameter diversity defined over an r-HST.

5. The Hierarchical Move Making Algorithm

In the first part of this section we propose a move making algorithm for the hierarchical P^n Potts model (defined in the previous section). In the second part, we show how our hierarchical move making algorithm can be used to address the general parsimonious labeling problem with optimality guarantees (strong multiplicative bound).

5.1. The Hierarchical Move Making Algorithm for the Hierarchical P^n Potts Model

In the hierarchical P^n Potts model the clique potentials are of the form of the diameter diversity defined over a given r-HST metric function. The move making algorithm proposed in this section to minimize such an energy function is a divide-and-conquer based approach, inspired by the work of [21]. Instead of solving the actual problem, we divide the problem into smaller subproblems where each subproblem is a P^n Potts model, which can be solved efficiently using $\alpha-$expansion [17]. More precisely, given an r-HST, each node of the r-HST corresponds to a subproblem. We start with the bottom node of the r-HST, which is a leaf node, and go up in the hierarchy solving each subproblem associated with the nodes encountered.

In more detail, consider a node p of the given r-HST. Recall that any node p in the r-HST is a cluster of labels denoted as $\mathcal{L}^p \subseteq \mathcal{L}$ (Figure 1). In other words, the leaf nodes of the subtree rooted at p belongs to the subset $\mathcal{L}^p$. Thus, the subproblem defined at node p is to find the labeling $\mathbf{x}^p$ where the label set is restricted to $\mathcal{L}^p$, as defined below.

$$\mathbf{x}^p = \operatorname*{argmin}_{\mathbf{x}\in(\mathcal{L}^p)^N} \Big(\sum_{i\in\mathcal{V}} \theta_i(x_i) + \sum_{c\in\mathcal{C}} w_c \delta^{dia}(\Gamma(\mathbf{x}_c)) \Big). \tag{7}$$

If p is the root node, then the above problem (equation (7)) is as difficult as the original labeling problem (since $\mathcal{L}^p = \mathcal{L}$). However, if p is the leaf node then the solution of the problem associated with p is trivial, $x_i^p = p$ for all $i \in \mathcal{V}$, that is, assign the label p to all the random variables. This insight leads to the design of our approximation algorithm, where we start by solving the simple problems corresponding to the leaf nodes, and use the labelings obtained to address the more difficult problem further up the

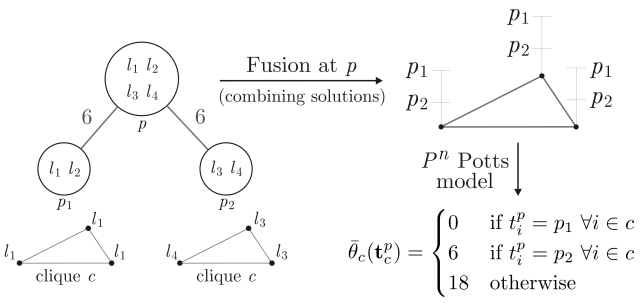


Figure 2: *An example of solving the labeling problem at non-leaf node (p) by combining the solutions of its child nodes $\{p_1, p_2\}$, given clique c and the labelings that it has obtained at the child nodes. The labeling fusion instance shown in this figure is the top two levels of the r-*HST *in the Figure 1. The diameter diversity of the labeling of clique c at node p_1 is 0 as it contains only one unique label l_1. The diameter diversity of the labeling at p_2 is $d^t(l_3, l_4) = 6$ and at p is $\max_{\{l_i,l_j\}\in\{l_1,l_3,l_4\}} d^t(l_i, l_j) = 18$.*

hierarchy. In what follows, we describe how the labeling of the problem associated with the node p, when p is a non-leaf node, is obtained using the labelings of its children node.

Solving the Parent Labeling Problem. Before delving into the details, let us define some notations for the purpose of clarity. Let T be the depth (or the number of levels) of the given r-HST and $\mathcal{N}(\tau)$ be the set of nodes at level τ. The root node is at the top level ($\tau = 1$). Let $\eta(p)$ denotes the set of child nodes associated with a non-leaf node p and $\eta(p, k)$ denotes its k^{th} child node. Recall that our approach is bottom up. Therefore, for each child node of p we already have an associated labeling. We denote the labeling associated with the k^{th} child of the node p as $\mathbf{x}^{\eta(p,k)}$. Thus, $x_i^{\eta(p,k)}$ denotes the label assigned to the i^{th} random variable by the labeling of the k^{th} child of the node p. We also define an N dimensional vector $\mathbf{t}^p \in \{1, \cdots, |\eta(p)|\}^N$, where $|\eta(p)|$ is the number of child nodes of node p. More precisely, for a given $\mathbf{t}^p$, $t_i^p = k$ denotes that the label for the i^{th} random variable comes from the k^{th} child of the node p. Therefore, the labeling problem at node p reduces to finding the optimal $\mathbf{t}^p$. Thus, the labeling problem at node p amounts to finding the best child index $k \in \{1, \cdots, |\eta(p)|\}$ for each random variable $i \in \mathcal{V}$ so that the label assigned to the random variable comes from the labeling of the k^{th} child (step 7, Algorithm 1).

Using the above notations, associated with a $\mathbf{t}^p$ we define a new energy function as:

$$E(\mathbf{t}^p) = \sum_{i\in\mathcal{V}} \bar{\theta}_i(t_i^p) + \sum_{c\in\mathcal{C}} w_c \bar{\theta}_c(\mathbf{t}_c^p). \tag{8}$$

where

$$\bar{\theta}_i(t_i^p) = \theta_i(x_i^{\eta(p,k)}) \quad \textit{if} \quad t_i^p = k. \tag{9}$$

Algorithm 1 The Move Making Algorithm for the Hierarchical P^n Potts Model.

input r-HST Metric; $w_c, \forall c \in \mathcal{C}$; and $\theta_i(x_i), \forall i \in \mathcal{V}$

1: $\tau = T$, the leaf nodes
2: **repeat**
3: **for** each $p \in \mathcal{N}(\tau)$ **do**
4: **if** $|\eta(p)| = 0$, leaf node **then**
5: $x_i^p = p, \forall i \in \mathcal{V}$
6: **else**
7: Fusion Move

$$\hat{\mathbf{t}}^p = \underset{\mathbf{t}^p\in\{1,\cdots,|\eta(p)|\}^N}{\operatorname{argmin}} E(\mathbf{t}^p) \tag{10}$$

8: $x_i^p = x_i^{\eta(p,\hat{t}_i^p)}$.
9: **end if**
10: **end for**
11: $\tau \leftarrow \tau - 1$
12: **until** $\tau > 0$.

In other words, the unary potential for $t_i^p = k$ is the unary potential associated to the i^{th} random variable corresponding to the label $x_i^{\eta(p,k)}$.

The new clique potential $\bar{\theta}_c(\mathbf{t}_c^p)$ is as defined below:

$$\bar{\theta}_c(\mathbf{t}_c^p) = \begin{cases} \gamma_k^p, & \text{if } t_i^p = k, \forall i \in c, \\ \gamma_{max}^p, & \text{otherwise}, \end{cases} \tag{11}$$

where $\gamma_k^p = \delta_{dia}(\Gamma(\mathbf{x}_c^{\eta(p,k)}))$ is the diameter diversity of the set of *unique labels* associated with $\mathbf{x}_c^{\eta(p,k)}$ and $\gamma_{max}^p = \delta_{dia}(\bar{\mathcal{L}}^p)$. The set $\bar{\mathcal{L}}^p = \cup_{k\in\eta(p)}\Gamma(\mathbf{x}_c^{\eta(p,k)})$ is the union of the unique labels associated with the child nodes of p. Recall that, because of the construction of the r-HST, $\mathcal{L}^q \subset \bar{\mathcal{L}}^p \subseteq \mathcal{L}^p$ for all $q \in \eta(p)$. Hence, the monotonicity property of the diameter diversity (Definition 1) ensures that $\gamma_{max}^p > \gamma_k^p, \forall k \in \eta(p)$. This is the sufficient criterion to prove that the potential function defined by equation (11) is a P^n Potts model. Therefore, the $\alpha-$expansion algorithm can be used to obtain the locally optimal $\mathbf{t}^p$ for the energy function (8). Given the locally optimal $\hat{\mathbf{t}}^p$, the labeling $\mathbf{x}^p$ at node p can be trivially obtained as follows: $x_i^p = x_i^{\eta(p,\hat{t}_i^p)}$. In other words, the final label of the i^{th} random variable is the one assigned to it by the $(\hat{t}_i^p)^{th}$ child of node p.

Figure 2 shows an instance of the above mentioned algorithm to combine the labelings of the child nodes to obtain the labeling of the parent node. The complete hierarchical move making algorithm for the hierarchical P^n Potts model is shown in the Algorithm 1.

Multiplicative Bound. Theorem 1 gives the multiplicative bound for the move making algorithm for the hierarchical P^n Potts model.

Algorithm 2 The Move Making Algorithm for the Parsimonious Labeling Problem.

input Diversity $(\mathcal{L}, \delta)$; $w_c, \forall c \in \mathcal{C}$; $\theta_i(x_i), \forall i \in \mathcal{V}$; $\mathcal{L}$; k

1: Approximate the given diversity as the mixture of k hierarchical P^n Potts model using Algorithm 3.
2: **for** each hierarchical P^n Potts model in the mixture **do**
3: Use the hierarchical move making algorithm defined in the Algorithm 1.
4: Compute the corresponding energy.
5: **end for**
6: Choose the solution with the minimum energy.

Algorithm 3 Diversity to Mixture of Hierarchical P^n Potts Model.

input Diversity $(\mathcal{L}, \delta)$, k

1: Compute the induced metric, $d(.,.)$, where $d(l_i, l_j) = \delta(\{l_i, l_j\}), \forall l_i, l_j \in \mathcal{L}$.
2: Approximate $d(.,.)$ into mixture of k r-HST metrics $d^t(.,.)$ using the algorithm proposed in [10].
3: **for** each r-HST metrics $d^t(.,.)$ **do**
4: Obtain the corresponding hierarchical P^n Potts model by defining the diameter diversity over $d^t(.,.)$
5: **end for**

Theorem 1. *The move making algorithm for the hierarchical P^n Potts model, Algorithm 1, gives the multiplicative bound of $\left(\frac{r}{r-1}\right)\min(\mathcal{M}, |\mathcal{L}|)$ with respect to the global minima. Here, $\mathcal{M}$ is the size of the largest maximal-clique and $|\mathcal{L}|$ is the number of labels.*

Proof Sketch. The factor of $\min(\mathcal{M}, |\mathcal{L}|)$ comes from the fact that each subproblem amounts to solving $\alpha-$expansion for the P^n Potts models. The factor of $\left(\frac{r}{r-1}\right)$ comes from the fact that the edge lengths of the r-HST forms a geometric progression (refer to Figure 1), therefore, the distance between any two leaf node is upperbounded by $e^{max}\left(\frac{r}{r-1}\right)$, where e^{max} is the length of the longest edge. Please see the technical report for the detailed proof. □

5.2. The Move Making Algorithm for the Parsimonious Labeling

In the previous subsection, we proposed a hierarchical move making algorithm for the hierarchical P^n Potts model (Algorithm 1). This restricted us to a small class of clique potentials. In this section we generalize our approach to the much more general parsimonious labeling problem.

The move making algorithm for the parsimonious labeling problem is shown in Algorithm 2. Given diversity based clique potentials, non-negative clique weights, and arbitrary unary potentials, Algorithm 2 approximates the diversity into a mixture of hierarchical P^n Potts models (using Algorithm 3) and then use the previously defined Algorithm 1 on each of the hierarchical P^n Potts models.

The algorithm for approximating a given diversity into a mixture of hierarchical P^n Potts models is shown in Algorithm 3. The first and the third steps of the Algorithm 3 have already been discussed in the previous sections. The second step, which amounts to finding the mixture of r-HST metrics for a given metric, can be solved using the randomized algorithm proposed in [10]. We refer the reader to [10] for further details of the algorithm for approximating a metric using a mixture of r-HST metrics.

Multiplicative Bound. Theorem 2 gives the multiplicative bound for the parsimonious labeling problem, when the clique potentials are any general diversity. Notice that the multiplicative bound of our algorithm is significantly better than the multiplicative bound of SoSPD [12], which is $\mathcal{M}\frac{\max_c \delta(\Gamma(\mathbf{x}_c)}{\min_c \delta(\Gamma(\mathbf{x}_c)}$.

Theorem 2. *The move making algorithm defined in Algorithm 2 gives the multiplicative bound of $\left(\frac{r}{r-1}\right)(|\mathcal{L}| - 1)O(\log|\mathcal{L}|)\min(\mathcal{M}, |\mathcal{L}|)$ for the* parsimonious labeling *problem (equation (4)). Here, $\mathcal{M}$ is the size of the largest maximal-clique and $|\mathcal{L}|$ is the number of labels.*

Proof Sketch. The additional factor of $(|\mathcal{L}| - 1)$ and $O(\log|\mathcal{L}|)$ comes from the inequalities $\delta(\mathcal{L}) \leq (|\mathcal{L}| - 1)\delta^{dia}(\mathcal{L})$ [4] and $d(.,.) \leq O(\log|\mathcal{L}|)d^t(.,.)$ [10], respectively. Technical report contains the detailed proof. □

Time Complexity. Each expansion move of our Algorithm 2 amounts to solving a graph-cut on a graph with $2|\mathcal{C}|$ auxiliary variables and $|\mathcal{C}|(2\mathcal{M}+2)$ edges (in worst case), therefore, the time complexity is $O((|\mathcal{V}| + |\mathcal{C}|)^2|\mathcal{C}|\mathcal{M})$. In addition, each subproblem in our algorithm is defined over a much smaller label set (number of child nodes). Furthermore, Algorithm 2 can be parallelized over the trees and over the subproblems at any level of the hierarchy. In contrast, each expansion move of SoSPD [12] amounts to solving submodular max-flow, which is $O(|\mathcal{V}|^2|\mathcal{C}|2^{\mathcal{M}})$ [11], exponential in the size of the largest clique. Furthermore, each expansion move of Ladicky *et al.* [22] amounts to solving a graph-cut on a graph with $|\mathcal{C}||\mathcal{L}|$ auxiliary nodes and $|\mathcal{C}|(2\mathcal{M} + |\mathcal{L}|)$ edges having a time complexity of $O((|\mathcal{V}| + |\mathcal{C}||\mathcal{L}|)^2|\mathcal{C}|(\mathcal{M} + |\mathcal{L}|))$ [13]. As can be seen from the above discussion, our move-making algorithm is significantly more efficient for the parsimonious labeling problem.

6. Experiments

We demonstrate the utility of parsimonious labeling on both synthetic and real data. In the case of synthetic data, we perform experiments on a large number of grid lattices and evaluate our method based on the energy and the time taken. We show the modeling capabilities of the parsimonious labeling by applying it on two challenging real problems: (i) stereo matching, and (ii) image inpainting. We use

the move-making algorithm for the co-occurrence based energy function proposed by Ladicky *et al.* [22] as our baseline. Based on the synthetic and the real data results, supported by the theoretical guarantees, we show that the move making algorithm proposed in our work outperforms [22].

Recall that the energy function we are interested in minimizing is defined in the equation (4). In our experiments, we frequently use the truncated linear metric. We define it below for the sake of completeness.

$$\theta_{i,j}(l_a, l_b) = \lambda \min(|l_a - l_b|, M), \forall l_a, l_b \in \mathcal{L}. \quad (12)$$

where λ is the weight associated with the metric and M is the truncation constant.

6.1. Synthetic Data

We consider following two cases: (i) given the hierarchical P^n Potts model, and (ii) given a general diversity based clique potential. In each of the two cases, we generate lattices of size 100×100, 20 labels, and use $\lambda = 1$. The cliques are generated using a window of size 10×10 in a sliding window fashion. The unary potentials are randomly sampled from the uniform distribution defined over the interval $[0, 100]$. In the first case, we randomly generate 100 lattices and random r-HST trees associated with each lattice, ensuring that they satisfy the properties of the r-HST. Each r-HST is then converted into hierarchical P^n Potts model by defining diameter diversity over each of them. The hierarchical P^n Potts model is then used as the actual clique potential. We performed 100 such experiments. On the other hand, in the second case, for a given value of the truncation M, we generate a truncated linear metric and 100 lattices. We treat this metric as the induced metric of a diameter diversity and apply Algorithm 1 for the energy minimization. We used four different values of the truncation factor $M \in \{1, 5, 10, 20\}$. For both the experiments, we used 7 different values of w_c: $w_c \in \{0, 1, 2, 3, 4, 5, 100\}$.

The average energy and the time taken for both the methods and both the cases are shown in the Figure 3. It is evident from the figures that our method outperforms [22] both in terms of time and the energy. In case (ii), despite the fact that our method first approximates the given diversity into mixture of hierarchical P^n Potts models, it outperforms [22]. This can be best supported by the fact that our algorithm has a strong multiplicative bound.

6.2. Real Data

In case of real data, the high-order cliques are the superpixels obtained using the mean-shift method [6], the clique potentials are the diameter diversity of the truncated linear metric (equation (12)). A truncated linear metric enforces smoothness in the pairwise setting, therefore, its diameter diversity will naturally enforce smoothness in the high-order cliques, which is a desired cue for the two applications we are dealing with. In both the real experiments we use $w_c = \exp^{-\frac{\rho(\mathbf{x}_c)}{\sigma^2}}$ (for high order cliques), where $\rho(\mathbf{x}_c)$ is the variance of the intensities of the pixels in the clique $\mathbf{x}_c$ and σ is a hyperparameter.

6.2.1 Stereo Matching

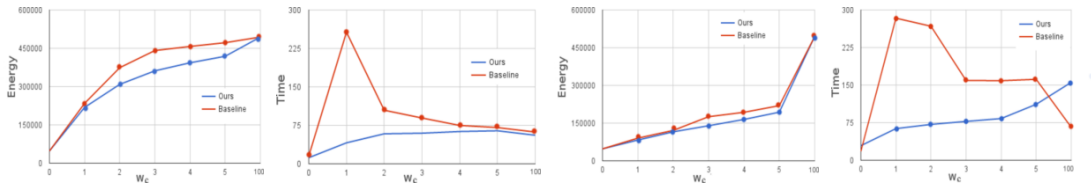
Given two rectified stereo pair of images, the problem of stereo matching is to find the disparity (gives the notion of depth) of each pixel in the reference image [26, 27]. We extend the standard setting of the stereo matching [26] to high-order cliques and test our method to the images, 'tsukuba' and 'teddy', from the widely used Middlebury stereo data set [25]. The unaries are the $L1-$norm of the difference in the RGB values of the left and the right image pixels. Notice that the index for the right image pixel is the index for the left image pixel minus the disparity, where disparity is the label. For 'tsukuba' and 'teddy' we used 16 and 60 labels respectively. In case of 'teddy' the unaries are trucated at 16. The weights w_c for the pairwise cliques are set to be proportional to the $L1-$norm of the gradient of the intensities of the neighboring pixels $\|\nabla\|_1$. In case of 'tsukuba', if $\|\nabla\|_1 < 8$, $w_c = 2$, otherwise $w_c = 1$. In case of 'teddy', if $\|\nabla\|_1 < 10$, $w_c = 3$, otherwise $w_c = 1$. As mentioned earlier, w_c for the high-order cliques is set to be proportional to the variance. We used different values of σ, λ, and M. Because of the space constraints we are showing results for the following setting: for 'tsukuba', $\lambda = 20$, $\sigma = 100$, and $M = 5$; for 'teddy', $\lambda = 20$, $\sigma = 100$, and $M = 1$. Figure 4 shows the results obtained. Notice that our method significantly outperforms [22] in terms of energy and the visual quality for both 'tsukuba' and 'teddy'.

6.2.2 Image Inpainting and Denoising

Given an image with added noise and obscured regions (regions with missing pixels), the problem is to denoise the image and fill the obscured regions such that it is consistent with the surroundings. We perform this experiment on the images, 'penguin' and 'house', from the widely used Middlebury data set. The images under consideration are gray scale, therefore, there are 256 labels in the interval $[0, 255]$, each representing an intensity value. The unaries for each pixel (or node) corresponding to a particular label is the squared difference between the label and the intensity value at that pixel. The weights w_c for the pairwise cliques are all set to one. We used different values of σ, λ, and the truncation M. Because of the space constraints we are showing results for the following setting: 'penguin', the $\lambda = 40$, $\sigma = 10000$ and $M = 40$; for 'house', the $\lambda = 50$, $\sigma = 1000$ and $M = 50$. Notice that our method (Figure 5) significantly outperforms [22] in terms of energy and visual quality for both 'penguin' and 'house'.

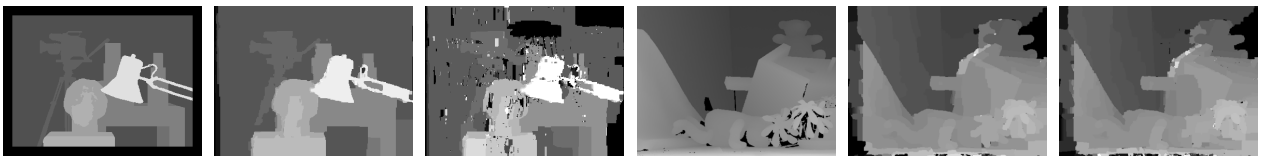
7. Discussion

We proposed a new family of labeling problems, called parsimonious labeling, where the energy function is defined using a diversity measure. Our energy function includes the



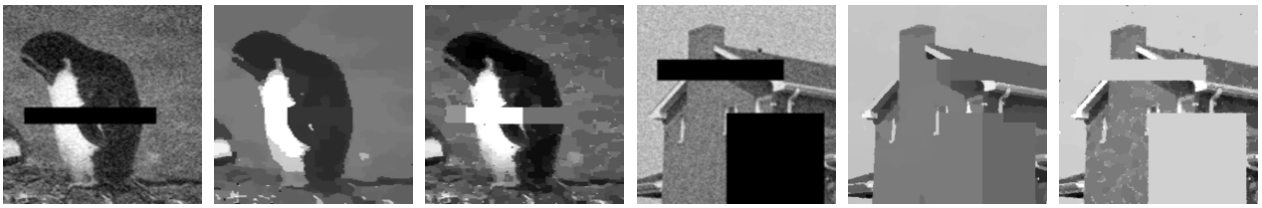
(a) Energy (b) Time (in seconds) (c) Energy (d) Time (in seconds)

Figure 3: *Synthetic (Blue: Our, Red: Co-occ [22]). The x-axis of all the figures is the weight associated with the cliques* (w_c). *Figures (a) and (b) are the plots when the hierarchical* P^n *Potts model is known. Figures (c) and (d) are the plots when a diversity (diameter diversity over truncated linear metric) is given as the clique potentials which is then approximated using the mixture of hierarchical* P^n *Potts model. Notice that in both the cases our method outperforms the baseline [22] both in terms of energy and time. Also, for very high value of* $w_c = 100$, *both the methods converges to the same labeling. This is expected as a very high value of* w_c *enforces rigid smoothness by assigning everything to the same label.*



(a) Tsukuba (Energy, Time (sec)) (b) Our ($\mathbf{1387300}$, 773) (c) Co-occ [22] (2117804, $\mathbf{306}$) (d) Teddy (Energy, Time (sec)) (e) Our ($\mathbf{3196400}$, 2918) (f) Co-occ [22] (3256200, $\mathbf{1392}$)

Figure 4: *Stereo Matching Results. Figures (a) and (d) are the ground truth disparity for the 'tsukuba' and 'teddy' respectively. Our method significantly outperforms the baseline Co-ooc [22] in both the cases in terms of energy. Our results are visually more appealing also. Figures (b) and (e) clearly shows the influence of 'parsimonious labeling' as the regions are smooth and the discontinuity is preserved. Recall that we use super-pixels obtained using the mean-shift as the cliques.*



(a) Penguin (Energy, Time (sec)) (b) Our ($\mathbf{12547000}$, 1964) (c) Co-occ [22] (14710302, $\mathbf{358}$) (d) House (Energy, Time (sec)) (e) Our ($\mathbf{36840076}$, 2838) (f) Co-occ [22] (41980784, $\mathbf{912}$)

Figure 5: *Image inpainting results. Figures (a) and (d) are the input images of 'penguin' and 'house' with added noise and obscured regions. Our method, (b) and (e), significantly outperforms the baseline [22] in both the cases in terms of energy. Visually, our method gives much more appealing results. We use super-pixels obtained using the mean-shift as the cliques.*

novel hierarchical P^n Potts model, which allows us to design an efficient and accurate move-making algorithm based on iteratively solving the minimum st-cut problem.

The large class of energy functions covered by parsimonious labeling can be used for various computer vision tasks such as semantic segmentation (where diversity function can be used to favor certain subsets of semantic classes), or 3D reconstruction (where diversity function can be used to favor certain subsets of depth values).

An interesting direction for future research would be to explore different diversities and propose specific algorithms for them, which may provide better theoretical guarantees. Another interesting work would be to directly approximate diversities into a mixture of hierarchical P^n Potts model, without the use of the intermediate r-HST metrics. This may also lead to better multiplicative bounds.

Acknowledgments. Puneet Dokania is partially funded by the European Community's Seventh Framework Programme under the ERC Grant agreement number 259112 and MOBOT Grant agreement number 600796, and Pôle de Compétitivité Medicen/ADOC Grant agreement number l11012185.

References

[1] Y. Bartal. On approximating arbitrary metrics by tree metrics. In *STOC*, 1998.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In *PAMI*, 2001.

[3] D. Bryant and P. F. Tupper. Hyperconvexity and tight-span theory for diversities. In *Advances in Mathematics*, 2012.

[4] D. Bryant and P. F. Tupper. Diversities and the geometry of hypergraphs. In *Discrete Mathematics and Theoretical Computer Science*, 2014.

[5] C. Chekuri, S. Khanna, J. Naor, and L. Zosin. Approximation algorithms for the metric labeling problem via a new linear programming formulation. In *SODA*, 2001.

[6] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. In *PAMI*, 2002.

[7] A. Delong, L. Gorelick, O. Veksler, and Y. Boykov. Minimizing energies with hierarchical costs. In *IJCV*, 2012.

[8] A. Delong, A. Osokin, H. Isack, and Y. Boykov. Fast approximate energy minimization with label costs. In *CVPR*, 2010.

[9] N. El-Zehiry and L. Grady. Fast global optimization of curvature. In *CVPR*, 2010.

[10] J. Fakcharoenphol, S. Rao, and K. Talwar. A tight bound on approximating arbitrary metrics by tree metrics. In *STOC*, 2003.

[11] A. Fix, T. Joachims, S. M. Park, and R. Zabih. Structured learning of sum-of-submodular higher order energy functions. In *ICCV*, 2013.

[12] A. Fix, C. Wang, and R. Zabih. A primal-dual algorithm for higher-order multilabel markov random fields. In *CVPR*, 2014.

[13] A. V. Goldberg, S. Hed, H. Kaplan, R. E. Tarjan, and R. F. Werneck. Maximum flows by incremental breadth-first search. In *European Symposium on Algorithms*, 2011.

[14] S. Gould, F. Amat, and D. Koller. Alphabet soup: A framework for approximate energy minimization. In *CVPR*, 2009.

[15] H. Y. Jung, K. M. Lee, and S. U. Lee. Stereo reconstruction using high order likelihood. In *ICCV*, 2011.

[16] J. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. In *FOCS*, 1999.

[17] P. Kohli, M. Kumar, and P. Torr. P3 & beyond: Solving energies with higher order cliques. In *CVPR*, 2007.

[18] P. Kohli, L. Ladicky, and P. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008.

[19] V. Kolmogorov. Minimizing a sum of submodular functions. In *Discrete Applied Mathematics*, 2012.

[20] M. P. Kumar. Rounding-based moves for metric labeling. In *NIPS*, 2014.

[21] M. P. Kumar and D. Koller. MAP estimation of semi-metric MRFs via hierarchical graph cuts. In *UAI*, 2009.

[22] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Graph cut based inference with co-occurrence statistics. In *ECCV*, 2010.

[23] X. Lan, S. Roth, D. Huttenlocher, and M. Black. Efficient belief propagation with learned higher-order markov random fields. In *ECCV*, 2006.

[24] R. B. Potts. Some generalized order-disorder transformations. In *Cambridge Philosophical Society*, 1952.

[25] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *IJCV*, 2002.

[26] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. In *PAMI*, 2008.

[27] M. Tappen and W. Freeman. Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *ICCV*, 2003.

[28] D. Tarlow, I. Givoni, and R. Zemel. Hop-map: Efficient message passing with high order potentials. In *AISTATS*, 2010.

[29] O. Veksler. Efficient graph-based energy minimization methods in computer vision, 1999.

[30] S. Vicente, V. Kolmogorov, and C. Rother. Joint optimization of segmentation and appearance models. In *ICCV*, 2009.

[31] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon. Global stereo reconstruction under second order smoothness priors. In *CVPR*, 2008.