Hierarchical CRF with Product Label Spaces for Parts-based Models

Gemma Roig[1,*] Xavier Boix[2,*]
Fernando De la Torre[3] Joan Serrat[2] Carles Vilella[1]

[1]La Salle, Universitat Ramon Llull. Barcelona 08022, Spain.

[2]Centre de Visió per Computador. Universitat Autònoma de Barcelona, 08193, Spain.

[3] Robotics Institute, Carnegie Mellon University. Pittsburgh, PA 15213, USA.

***Abstract*— Non-rigid object detection is a challenging open research problem in computer vision. It is a critical part in many applications such as image search, surveillance, human-computer interaction or image auto-annotation. Most successful approaches to non-rigid object detection make use of part-based models. In particular, Conditional Random Fields (CRF) have been successfully embedded into a discriminative parts-based model framework due to its effectiveness for learning and inference (usually based on a tree structure). However, CRF-based approaches do not incorporate global constraints and only model pairwise interactions. This is especially important when modeling object classes that may have complex parts interactions (e.g. facial features or body articulations), because neglecting them yields an oversimplified model with suboptimal performance. To overcome this limitation, this paper proposes a novel hierarchical CRF (HCRF). The main contribution is to build a hierarchy of part combinations by extending the label set to a hierarchy of product label spaces. In order to keep the inference computation tractable, we propose an effective method to reduce the new label set. We test our method on two applications: facial feature detection on the Multi-PIE database and human pose estimation on the Buffy dataset.**

I. Introduction

The seminal work by Fischler and Elschlager [10] proposed a parts-based model as the parametrization of an object class with a set of parts able to represent its shape or structure. Each part has to be consistent with different instances of the same object, and it corresponds to significant locations of the object such as boundaries or distinguished landmarks. This representation addresses a central problem in computer vision: the localization of the object parts in an image.

One of the most successful approaches for parts-based models builds upon a set of candidates that potentially correspond to an object part, and it selects the candidates that jointly better fit the parts-based model [7]. Thus, it aims to assign to each part a label that represents a candidate. State-of-the-art methods that evaluate such labeling use Conditional Random Fields (CRFs) as an energy function [1], [9], [21], that naturally relates parts with their relative location variability. Typically, these relations are defined between pairs of parts, and the matching cost is based on an individual appearance models for each part. Even though these models lead to tractable inference, they are unable to capture important patterns between sets of more than two parts, which are present in most object classes. Such global patterns appear when detecting faces or bodies because, for instance, they are mostly symmetric or the skin color of a person is mostly constant. Neglecting this global structure

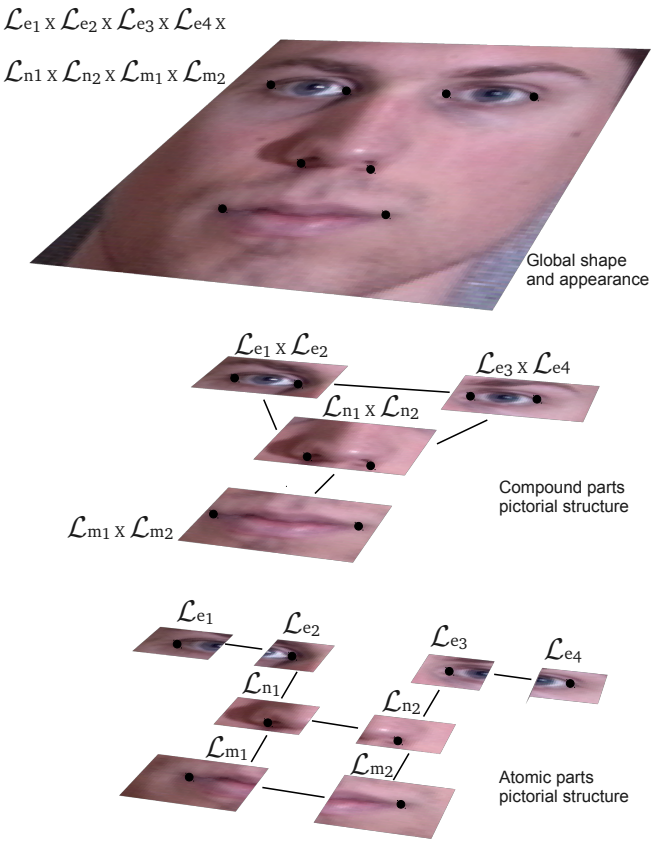

Fig. 1. HCRF of parts for facial feature detection. Each part, simple or compound, represents a node. *Connections between levels are not drawn to simplify the figure.* Atomic parts are at the lowest level and hierarchically compose the parts of the higher levels. The label sets form a hierarchy of product label spaces, which are able to encode multiple atomic parts simultaneously and model higher-order relations.

*Both first authors contributed equally. They are currently affiliated at ETH Zurich, Switzerland.

This work was partially supported by the Spanish Ministry of Science and Innovation under projects TRA2010-21371-C03-01 and Consolider Ingenio 2010 MIPRCV (CSD200700018). Gemma Roig acknowledges the support of the University, Research and Information Society Department of Catalonia Government, and Xavier Boix the FPU fellowship AP2008-03378. Fernando De la Torre was partially supported by the National Institute of Health Grant R01 MH 051435.

might yield to myopic models. To address this problem, this paper proposes a hierarchical conditional random field model (HCRF) that provides a richer structure without sacrificing inference effectiveness.

Lan and Huttenlocher [18] previously noticed the importance of including complex shape relations, proposing high-order cliques to model relative positions involving several parts. Caetano *et al.* [2], [20] introduced several graphs able to impose a global shape constraint without using a fully connected graph, and [14] proposed a method that learns the most meaningful connections that preserve the shape of the object class. These graphs yield an effective inference, though the global constraints are purely geometric (*e.g.* rigidity of the whole graph) and are not adequate for modeling appearance relations between parts. In order to include richer appearance relations in the context of human pose estimation, [22] extended the pairwise constraints to capture symmetry of clothing and smooth contour connections, and [6] considered appearance consistency between pairs of correlated parts. Recently, Sapp *et al.* showed promising results enabling richer appearance models for the pairwise constraints, either with a non-parametric method [25] or including contour continuation and segmentation cues [26]. Although these models use richer cues than previous ones, they are not defined to model sets of parts, and hence still suffer from restricted expressiveness. Our aim is to go one step further: model higher-order appearance and shape relationships.

Similar in spirit, [29] proposed a hierarchical CRF (HCRF) in which different levels correspond to a different granularity of the object class structure. The local level (the bottom) represents the atomic parts of the object class. The global level (the top) represents the whole object, and in between there are mid levels, which correspond to a set of multiple parts of the object. For example, if we consider a face as the object class, the local level could correspond to the corners of the eyes, the corners of the mouth and the nostrils. The mid level could represent the eyes, the mouth and the nose, and the global level might be the whole face. This approach takes into account sets of parts, and hence can use more adequate cues at each level. However, to make it tractable, [29] oversimplified the representation of the part candidates using simple labels, which summarize the set of combined parts. This might be valid for the lower levels closer to the atomic parts, but it does not model very well the higher levels. At these levels, far away from the atomic parts, it imposes a rather simplistic model because it encodes multiple parts together in a single label space.

This paper presents a novel HCRF structure that uses labels of multiple dimensions to represent each set of parts in the hierarchy. Instead of expressing all combined parts simplifying its representation, we preserve the original representation of all atomic parts using its product of label spaces, see Figure 1. In order to make the underlying optimization problem feasible, we introduce a branch-and-bound strategy to reduce the prohibitive cardinality of the product label space. Our model is able to integrate different cues adequate to describe each part in the hierarchy. In addition, it enables us to include richer relations both among and within sets of parts. To show this capability, we include in the HCRF a global shape model of a set of parts, which we learn with Principal Component Analysis (PCA), and an appearance model that takes into account color correlations among multiple parts. Our contribution is three-fold: *(i)* the hierarchy of product of label spaces, *(ii)* a strategy to reduce the cardinality of the product of label spaces, *(iii)* a global shape and appearance model embedded in the HCRF. Our method achieves state-of-the-art results in two challenging databases: Buffy [8] for human bodies and Multi-PIE [13] for faces.

II. Random Fields for Parts-based Models

This section introduces the CRFs formulation for the parts-based model. This formulation defines one random variable for each part and a set of candidate parts indexed with the label set $\mathcal{L} = \{l_1, l_2, \ldots, l_m\}$. Each random variable associates a part with one of the candidates taking a value from $\mathcal{L}$. The probability density function of how likely is to assign certain candidates to the parts is modeled with a CRF, and it can be represented with a graph $\mathcal{G} = (\mathcal{V}, \mathcal{N})$. The set $\mathcal{V}$ indexes the nodes that correspond to random variables, and $\mathcal{N} = \{\mathcal{N}_i\}$ is the neighborhood system of the random field, where $\mathcal{N}_i$ represents the set of all neighbors of node $i \in \mathcal{V}$. We use $\mathbf{X} = \{X_i\}$ to denote the set of random variables or nodes, and $\mathbf{x} = \{x_i\}$ a possible state or instantiation of $\mathbf{X}$.

Let $\mathcal{C}$ be the set of all maximal cliques[1] of the CRF. Then, the posterior $P(\mathbf{X} = \mathbf{x}|\mathbf{O})$ can be expressed as a Gibbs distribution with energy $E(\mathbf{x}) = \sum_{c \in \mathcal{C}} \varphi_c(\mathbf{x}_c, \mathbf{O})$ [15], where φ_c is the potential function of the maximal clique $c \in \mathcal{C}$, and $\mathbf{O}$ some observations or measurements. From now on, we omit the dependency of the potentials on $\mathbf{O}$ for notation simplicity. The most probable state $\mathbf{x}^*$ that maximizes the posterior probability (MAP) is

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} P(\mathbf{X} = \mathbf{x}|\mathbf{O}) = \arg\min_{\mathbf{x}} E(\mathbf{x}). \qquad (1)$$

Typically, most authors base the energy function on some pairwise shape restrictions and a matching score based on the individual appearance of each part. This can be written as the sum of the unary and consistency potentials:

$$\sum_{i \in \mathcal{V}} \phi_i(x_i) + \sum_{i \in \mathcal{V}, j \in \mathcal{N}_i} \psi_{ij}(x_i, x_j), \qquad (2)$$

where $\mathcal{N}_i$ is defined by pairwise relationships. The unary term $\phi_i(x_i)$ is based on a matching score that relies on some appearance or texture descriptors $\mathbf{O}_i$, where $\mathbf{O}_i$ is the observation that only affects X_i in the model. The consistency potential $\psi_{ij}(x_i, x_j)$ determines the cost to set labels x_i and x_j to the random variables X_i and X_j. Usually, its purpose is to penalize deviation from some shape constraint like proximity.

[1] A clique is a subgraph in which every node is connected to all other nodes in the subgraph, and it is maximal when it is not a subset on any other clique.

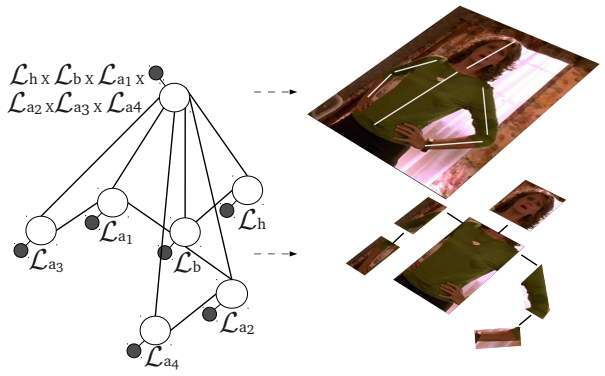

Fig. 2. HCRF for human pose estimation. At the top level we encode all parts using a hyperlabel space, which enables us to use a global shape and appearance model of the whole human body.

The pairwise CRF suffers from an inability to express dependency among several parts at the same time. It only exploits local information, while in most object classes there exist certain dependency among multiple parts. Despite this drawback, most authors adopt this framework and focus on integrating richer cues to the model [6], [8], [9], [21], [22], [25], [26]. In contrast, we propose Hierarchical CRF (HCRF), a model that is able to capture the dependencies among the parts.

III. HCRF with Hierarchy of Product Label Spaces

We now introduce a novel structure of HCRF able to model dependencies over sets of parts. The HCRF extends the basic CRF by including parts at different granularities. These parts are obtained as a combination of the original (atomic) parts, and thus, successively model a higher level part of the object class. Every new combination of parts is defined hierarchically. For example, in a human body we first could group the upper and forearm to form the whole arm, then the arm with the body, and so on. Fitting the whole hierarchy of parts is a complex task, but it allows us to include more reliable cues based on larger regions.

As in [29], for each part, simple or compound, we designate one random variable or node placed in the corresponding hierarchical graph. We add a superindex in our notation to refer to the level in the hierarchy. The set $\mathcal{V}^{(k)}$ are indexes that associate nodes at level k with their corresponding random variable, where $\mathcal{V}^{(k)} \subset \mathcal{V}$. Analogously, we define $\mathcal{N}_i^{(k)}$ as the set of neighbors at all levels of random variable $i \in \mathcal{V}^{(k)}$. Note that since it is a hierarchy, the neighbors of a k-level node, $\mathcal{N}_i^{(k)}$, can be at levels $(k-1)$, k or $(k+1)$. At the lowest level we place one node for each atomic part. Nodes in the immediately upper level are pairwise connected to nodes of the first level according to the hierarchy of parts. We continue in the same way for higher levels.

In [29], the compound parts are represented by summarizing all encoded parts with its central position. This summarization stands for keeping the same form of the label set $\mathcal{L}$ at all levels of the hierarchy, which yields tractable inference. Although this representation might be valid for lower levels, it does not model well higher levels. At these levels, where multiple atomic parts are related, it leads to a rather oversimplified model: it reduces the representation of all encoded parts as if they were one single part. In contrast, we propose to use multidimensional labels to take into account multiple parts at the same time. We represent a set of parts by the product of label spaces of the encoded atomic parts.

Let $\mathcal{L}^s = \mathcal{L} \times \mathcal{L} \times \ldots \times \mathcal{L}$ be the product label space built using s label sets $\mathcal{L}$. We refer as *hyperlabel* a label in $\mathcal{L}^s$, which is able to represent s atomic parts at the same time. Note that the cardinality of the set of hyperlabels is $|\mathcal{L}^s| = m^s$, where $m = |\mathcal{L}|$. Although this can be substantially large, in Section V we introduce an strategy that effectively reduces the size of the hyperlabel set under consideration. Let $\ell^a \in \mathcal{L}$ be the label placed at dimension $a < s$ of $\boldsymbol{\ell} \in \mathcal{L}^s$, *i.e.*

$$\boldsymbol{\ell} = (\ell^1, \ell^2, \ldots, \ell^a, \ldots, \ell^s). \tag{3}$$

At each dimension we place a label that corresponds to a candidate of an atomic part. *From now on, we redefine the label set of all nodes $i \in \mathcal{V}$ as $\mathcal{L}_i^{s_i}$*, where s_i is the number of parts encoded by the node. Each node has a different label set $\mathcal{L}_i^{s_i}$, which is determined by s_i and the atomic part that corresponds to each dimension. In the hierarchy, the hyperlabel of a k-level node encodes the hyperlabels of connected nodes at level $(k-1)$, reusing the label spaces. For instance, let $i \in \mathcal{V}^{(k)}$ be a k-level node and $j_1, j_2, \ldots \in \mathcal{N}_i^{(k)}$ all its connected nodes at level $(k-1)$, $j_1, j_2, \ldots \in \mathcal{V}^{(k-1)}$. The label set of node $i \in \mathcal{V}^{(k)}$ is

$$\mathcal{L}_i^{s_i} = \mathcal{L}_{j_1}^{s_{j_1}} \times \mathcal{L}_{j_2}^{s_{j_2}} \times \ldots \tag{4}$$

That is, the label set of a k-level node $\mathcal{L}_i^{s_i}$ is the product of label spaces of the connected $(k-1)$-level nodes. All label sets form a hierarchy of product label spaces following the hierarchy of parts. Figure 2 shows an example of how all label sets form a hierarchy.

The energy function of graph $\mathcal{G}$ is now the sum of the potentials at all levels,

$$\sum_k \left(\sum_{i \in \mathcal{V}^{(k)}} \phi_i(\mathbf{x}_i) + \sum_{i \in \mathcal{V}^{(k)}, j \in \mathcal{N}_i^{(k)}} \psi_{ij}(\mathbf{x}_i, \mathbf{x}_j) \right). \tag{5}$$

Recall that $\mathbf{x}_i$ is in boldface because it is a multidimensional vector of labels, and both potentials $\phi_i(\mathbf{x}_i)$ and $\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ relate sets of parts. The expressive power of our model stems from the capacity of expanded representation of the hyperlabels, which enables us to consider relations among multiple parts simultaneously. Since in our HCRF the same atomic part is encoded at different levels, we need a consistency potential between levels. This potential enforces a consistent labeling of the same atomic part in the hierarchy. In the next section, we define this potential and we also give several examples of other potentials that benefit from our model.

IV. Hyperlabel Potentials

Once we have presented the HCRF hyperlabels structure, we introduce several examples of potentials that are not supported by all previous methods based on HCRF. In this section, we distinguish the consistency potentials depending on whether the nodes are in the same or different levels.

A. Consistency between Levels

An important issue that has to be addressed is the possible inconsistency between nodes that encode the same atomic part. To overcome this issue the consistency potential enforces a close position of the same atomic part encoded at different nodes. For each pair of connected nodes $i, j \in \mathcal{V}$ at different levels, where $\mathbf{x}_i \in \mathcal{L}_i^{si}$ and $\mathbf{x}_j \in \mathcal{L}_j^{s_j}$, we define the set $\mathcal{P}_{ij}$ which contains all pairs of hyperlabel dimension indexes $(p, q) \in \mathcal{P}_{ij}$ that refer to the same atomic part. The consistency potential between levels is

$$\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j) = \max_{(p,q)\in\mathcal{P}_{ij}} \left\{\psi'_{ij}(x_i^p, x_j^q)\right\}, \tag{6}$$

where ψ'_{ij} evaluates the squared Euclidean distance of the two candidates x_i^p and x_j^q, and the $\max$ operator gets the worst case over all pairs of parts in $\mathcal{P}_{ij}$. This enforces all pairs of atomic parts that are in both product label spaces to be close to each other.

B. Unary and Consistency within Levels

We introduce four different strategies to define the potentials that relate parts within the same level. Their implementation-specific details are provided in Sections VI and VII.

1) Global Shape Potential: Our HCRF structure enables us to define a potential to evaluate the shape of multiple parts from a shape model learned on a training set. Let $(\mathbf{u}, \mathbf{v})$ be the vector of image coordinates of all parts encoded in the hyperlabel $\boldsymbol{\ell} \in \mathcal{L}_i^{s_i}$. Before learning a shape model, we apply Procrustes analysis to remove rigid transformations (e.g., rotation, scale and translation). We use $(\mathbf{u}_R, \mathbf{v}_R)$ to denote aligned coordinates with the rotation matrix $\mathbf{R}$. Once the training set has been aligned, we learn the shape model by computing the Principal Component Analysis (PCA) of the aligned coordinates. The PCA subspace is defined with $(\mathbf{B}, \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is the mean shape of the object examples, and $\mathbf{B}$ is the matrix of modes, *i.e.* the collection of eigenvectors associated with higher eigenvalues of the PCA. $\mathbf{B}$ encodes the possible shape variations learned from the training set.

We define the potential as the minimum distance between the PCA subspace and the evaluated shape $(\mathbf{u}, \mathbf{v})$. Thus, if $(\tilde{\mathbf{u}}, \tilde{\mathbf{v}})$ is the shape generated by the PCA with minimum distance to $(\mathbf{u}, \mathbf{v})$, the potential becomes

$$\phi_i(\mathbf{x}_i) = \sum_p w_p \|(u_p, v_p) - (\tilde{u}_p, \tilde{v}_p)\|_2^2, \tag{7}$$

where w_p weights each part p. The computation of $(\tilde{\mathbf{u}}, \tilde{\mathbf{v}})$ is not trivial because the PCA model has been previously learned with Procrustes alignment. Since we have two unknowns, $\mathbf{R}$ and $(\tilde{\mathbf{u}}, \tilde{\mathbf{v}})$, we proceed iteratively in two steps until convergence: *(i)* compute $\mathbf{R}$, and *(ii)* obtain $(\tilde{\mathbf{u}}, \tilde{\mathbf{v}})$ by doing the inverse transformation of $(\tilde{\mathbf{u}}_R, \tilde{\mathbf{v}}_R)$, which is computed as

$$[\tilde{\mathbf{u}}_R^T\ \tilde{\mathbf{v}}_R^T\ ]^T = \underbrace{\boldsymbol{\mu} + \mathbf{B}\overbrace{\left(\mathbf{B}^T\left(\left[\mathbf{u}_R^T\ \mathbf{v}_R^T\right]^T - \boldsymbol{\mu}\right)\right)}^{\text{projection to PCA subspace}}}_{\text{back-projection to original space}}. \tag{8}$$

That is, we first project $(\mathbf{u}_R, \mathbf{v}_R)$ to the PCA subspace, and then, back-project it to the original space. We initialize $(\tilde{\mathbf{u}}_R, \tilde{\mathbf{v}}_R)$ with the mean of the PCA model. In all tested cases the algorithm converges in less than 20 iterations.

2) Robust Pairwise Shape Potentials: When the number of encoded parts in a hyperlabel is small, our global shape model is unable to learn the object shape with the PCA. In this case, the learned PCA subspace comprises almost all the space of possible shapes, and hence is too flexible to capture the object shape variability. To overcome this limitation, we introduce a simpler model that relates all possible pairs of atomic parts. We proceed in two steps. We first evaluate all pairs of parts in the hyperlabel, and then select the worst pair. The evaluation of each pair of parts is done computing a score with the same Gaussian model as in [25]. Then, instead of simply averaging all scores, we merge them by selecting the worst one. This second step adds robustness because enforces a consistent shape between all pairs of parts.

3) Global Appearance Potential: The atomic parts are usually too local, and consequently its descriptors are not so discriminative. Since the hierarchy includes parts at different granularities, we can adapt the descriptors of each part at its level. At higher levels, where parts represent larger regions with a richer context, we can use more reliable cues. For instance, we can define a unary potential for the detection of the whole face, which is far more accurate than the unary potential of a single corner of an eye.

4) Color Correlation Potential: We introduce a potential that evaluates the color correlation within parts encoded in a hyperlabel. We first compute a color histogram $\mathbf{h}_p$ for each encoded part p, and then, we compute the one versus one distance $r_{pq} = d(\mathbf{h}_p, \mathbf{h}_q)$, using the intersection histogram distance [19]. The resulting feature vector $\mathbf{r}$ expresses the color similarity among encoded parts. As usual, the potential is based on a classification score obtained from $\mathbf{r}$.

V. Inference

We have presented a novel HCRF structure able to express relations among multiple parts. However, because it is built using hyperlabels, it suffers from an excessive cardinality of the label set, $|\mathcal{L}_i^{s_i}| = m^{s_i}$, which renders the inference in practice infeasible. Traditionally, for discrete probabilistic models with variables with very large domains, inference is achieved by reducing the label set by either discarding labels [11], [12] or sampling the label space [16], [28]. Analogously, we first use a branch-and-bound algorithm that prunes $\mathcal{L}_i^{s_i}$ selecting hyperlabels, and then, apply any suitable inference algorithm like Loopy Belief Propagation

(LBP) [17]. In this section, we focus on the selection of the hyperlabels, which enables us to effectively compute LBP.

Let $\mathbf{x}^*$ be the (unknown) inferred solution when using the large label set $\mathcal{L}_i^{s_i}$ at all nodes. We do not have access to $\mathbf{x}^*$ due to infeasible inference, but we might have a good approximation if inference is done over an equivalent smaller label set. We reduce the label set $\mathcal{L}_i^{s_i}$ of each node $i \in \mathcal{V}$, starting from the lowest level to the highest because our HCRF hierarchically reuses the label spaces. For each node, we select the $m' < m^{s_i}$ hyperlabels $\boldsymbol{\ell} \in \mathcal{L}_i^{s_i}$ with higher probability $P(\boldsymbol{\ell} = \mathbf{x}_i^*|\mathbf{O})$. This posterior models whether hyperlabel $\boldsymbol{\ell}$ is the non-approximated inferred solution $\mathbf{x}_i^*$ or not. It establishes a probability on the hyperlabel set, which in turn allows us to rank the most likely hyperlabels. It can be efficiently computed with the approximation (see the Appendix):

$$P(\boldsymbol{\ell} = \mathbf{x}_i^*|\mathbf{O}) \propto P(\boldsymbol{\ell}) \prod_{p<s_i} P(X_i^p = \ell^p|\mathbf{O}_i^p). \qquad (9)$$

The prior $P(\boldsymbol{\ell})$ models any joint shape constraint between the parts encoded in $\boldsymbol{\ell}$, and $P(X_i^p = \ell^p|\mathbf{O}_i^p)$ is the probability that the part encoded in dimension p takes a certain label ℓ^p, based on an observation $\mathbf{O}_i^p$ relative to this single part.

A branch-and-bound algorithm is able to *exhaustively* search among the whole space $\mathcal{L}_i^{s_i}$ by massively discarding large sets of fruitless hyperlabels. It establishes a search tree, where hyperlabels are build incrementally by increasing the number of encoded parts. At each level of the tree, it adds a part until it reaches the leafs, where the hyperlabels are. For instance, let $\boldsymbol{\ell}'' \in \mathcal{L}_i^n$ be a partially build hyperlabel at the n-th level of the search tree, and $\boldsymbol{\ell}' \in \mathcal{L}_i^{n+1}$ be equal to $\boldsymbol{\ell}''$ after a branching to the $(n+1)$-th level. Since branching is done by increasing the number of encoded atomic parts, we add an extra dimension to $\boldsymbol{\ell}''$ to build $\boldsymbol{\ell}'$. At the leafs of the search tree we obtain the hyperlabels in $\mathcal{L}_i^{s_i}$.

During the exploration of the tree, the algorithm maintains a set $\mathcal{S}$ of the $m' < m^{s_i}$ hyperlabels with the highest posterior $P(\boldsymbol{\ell} = \mathbf{x}_i^*|\mathbf{O})$. An upper bound of this posterior is evaluated for each partially build hyperlabel $\boldsymbol{\ell}' \in \mathcal{L}_i^n$. If the upper bound is lower than all the posteriors of the hyperlabels in the set $\mathcal{S}$, we can discard all hyperlabels build from $\boldsymbol{\ell}'$. Since these hyperlabels have a posterior lower or equal than the upper bound, we are sure that none of them has a posterior high enough to be selected. This pruning is what preserves a tractable computational cost. In our case, we define the upper bound of the posterior as

$$\gamma_{\boldsymbol{\ell}'} = P(\boldsymbol{\ell}') \prod_{p<n} P(X_i^p = \ell^p|\mathbf{O}_i^p). \qquad (10)$$

It is straightforward to check that this is in fact an upper bound if $P(\boldsymbol{\ell}') \geq P(\boldsymbol{\ell})$. The definition of the prior $P(\boldsymbol{\ell})$ must satisfy this condition. In Algorithm 1 we summarize a recursive implementation of our branch-and-bound.

```
function S=Branch&Bound(ℓ'', S, n)
    foreach l ∈ L do
        ℓ' = (ℓ'', l) ;                                  // Branch
        if ∃ℓ ∈ S : γ_ℓ' ≥ P(ℓ = x_i^*|O) then           // Bound
            if n = s_i then                              // If it is a leaf
                ℓ' ↣ S;                                  // Replace worst in S
            else
                S=Branch&Bound(ℓ', S, n + 1);
            end
        end
    end
end
```

Algorithm 1. Branch-and-bound algorithm for selecting the m' hyperlabels $\boldsymbol{\ell} \in \mathcal{L}_i^{s_i}$ with higher posterior $P(\boldsymbol{\ell} = \mathbf{x}_i^*|\mathbf{O})$. The set $\mathcal{S}$ stores the best found hyperlabels.

VI. Facial Feature Detection

This section reports experimental results on the problem of facial feature detection using the CMU Multi-PIE dataset [13]. It contains a total of 337 subjects, under 15 different viewpoints, 6 expressions and 19 illuminations. A total of $108,566$ images have been manually labeled with ground-truth of the 68 landmarks of the human face. We used $1,704$ images of frontal faces showing different expressions, and built a model to recover the location of 8 landmarks: the corners of the eyes, the corners of the mouth and the nostrils.

A. Implementation

Structure of the hierarchy: To recover the 8 landmarks we use the HCRF with three levels illustrated in Figure 1. The 8 atomic parts are represented in the first level. The second level is built by combining pairs of these parts to form the mouth, nose and eyes. This allows us to include larger meaningful parts of the face, which can be described with more reliable cues. The node at third level stands for the whole shape of the face. It includes all atomic parts to jointly evaluate the global shape of the resulting face.

Candidate parts: We obtained candidate points using the Harris corner detector, that [23] showed good performance to detect the corners of the eyes, mouth and nostrils. We used around 1000 candidates for the first level. For the

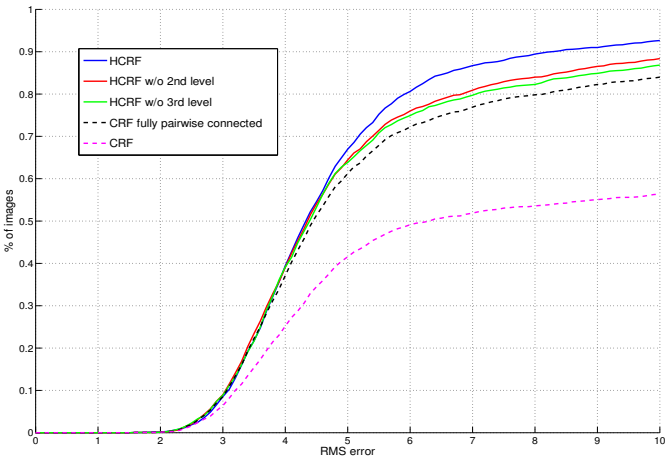

Fig. 3. Results on Multi-PIE. We evaluate our HCRF and the performance of each level. We compare our method with a CRF with pairwise potentials, using the connectivity pattern for the first level or fully connected.

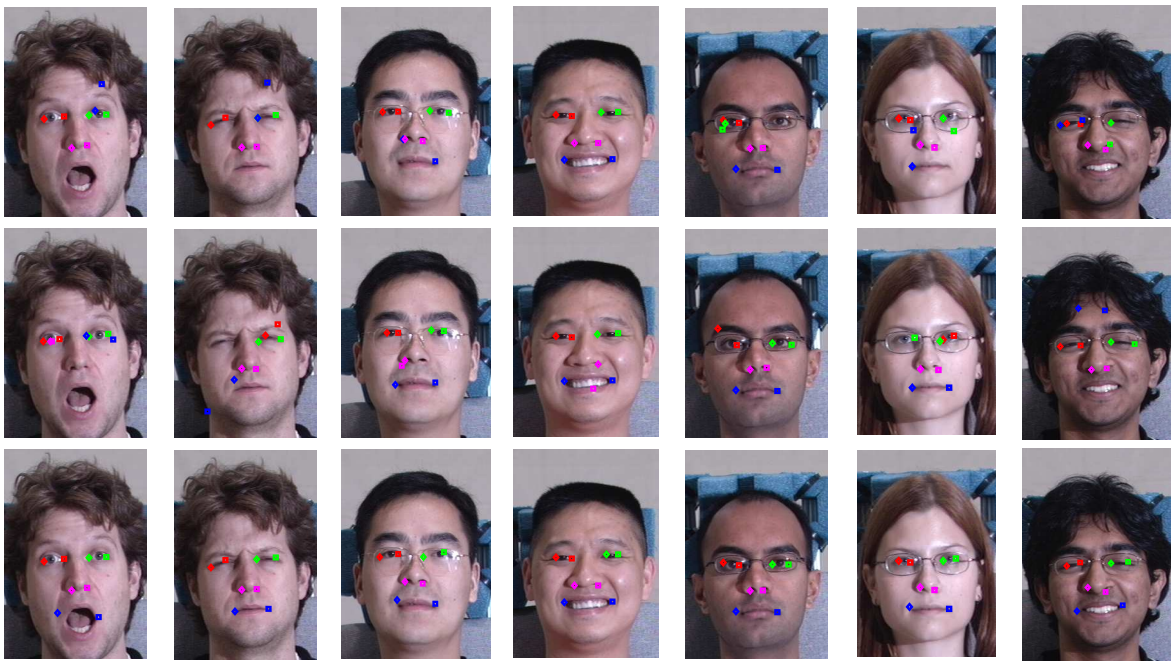

Fig. 4. Qualitative results on MultiPIE. We compare results of the CRF of the first level (top), the HCRF without the top level (middle) and the full HCRF (bottom). The eyes are indicated with red and green, the mouth with blue, and nostrils with magenta.

second level, we selected 1000 hyperlabels with the branch-and-bound strategy and 100 hyperlabels for the third level (without using the prior $P(\ell)$ of Eq. (9) by setting it always to 1).

Unary potentials: For the first level, we extracted a patch of 20×20 pixels around each candidate, and represented it using the SIFT descriptor. The second level models the global appearance for each part combination. To model all the region of the part, we concatenated the SIFT histograms of the encoded atomic parts. We used a SVM classifier with intersection kernel [19] to compute matching scores, trained with the same number of positive and negative samples. Positive examples were built with the Harris detected corners that were closer than 4 pixels to the ground-truth, and negative examples were randomly selected. For the third level, we used the global shape unary potential learned with PCA (only the first eigenmode).

Consistency Potentials within Levels: They are in the first and second levels. We defined each of them as robust pairwise shape potentials described in Section IV. The connections were set between parts closer to each other.

Learning HCRF parameters: We learned a weighting for each potential in the HCRF and also the Gaussian model of each robust pairwise shape potential. We discriminatively learned these parameters with a Gibbs-like sampling algorithm, which varies one single parameter at a time.

B. Experiments

We split the multi-PIE database into 4 subsets, such that the subject identities do not overlap. Then, we performed a 4-cross validation, where 3 sets were use for training and 1 for testing.

Evaluation of the HCRF: In Figure 3, we show the performance of the HCRF described above and also similar CRF-based methods. Results show that using a CRF with the connectivity pattern of our first level, the percentage of images with a root-mean-squared (RMS) error lower than 10 pixels is 57%, whereas with the three levels of hierarchy the score reaches to 93%. As we add the appearance of the parts at second level or the global shape model, the HCRF is able to capture more complex appearance and shape patterns. In Figure 4 results are depicted.

In order to show that high-order relations in fact do exist and that previous CRF-based methods are unable to capture them, we compare with a fully connected CRF with pairwise potentials. These potentials are the same Gaussian models as we use at the first level. Our HCRF clearly outperforms the fully connected CRF. This shows that our hyperlabels potentials are able to capture meaningful information not modeled by the CRF.

The HCRF of [29] is able to include parts at different granularities and describe them with adequate cues. Instead of using a product of label spaces it uses a simple label set that models each compound part with the central position of the encoded atomic parts. We evaluated our hierarchy using their label spaces, being the first and second levels equivalent. However, at the top node, all atomic parts are summarized with the central position of the face, and hence, we are unable to evaluate the global shape of the atomic parts. Thus, the representation of [29] is not suitable for the third level of our hierarchy.

Active Shape Models (ASM) [3] and extensions [4], [5], [24], [27] have been a popular tool for facial feature detection. These methods build holistic shape and appearance

models using a variant of PCA, and the search is typically based on deterministic optimization methods that may result in local minima. It is important to notice, that unlike our HCRF, the performance of these methods is rather dependent on the initialization. In the next section we will further illustrate the performance of our HCRF on the problem of human pose estimation.

VII. Human Pose Estimation

This section addresses the problem of 2D human pose estimation. This problem consists on localizing the body parts of a person in an image. To show the effectiveness of the HCRF we evaluated our method on the challenging Buffy dataset [8][2]. It is composed by frames from the TV show Buffy the Vampire Slayer. In each image a person is annotated with line segments indicating the position of its head, torso, upper arms and forearms.

A. Implementation

Structure of the Hierarchy: Figure 2 provides a two level hierarchy for human pose estimation. The highest level consists on a single node that encodes all atomic parts. It enables us to capture the color correlation among parts and the global shape. We omit any possible intermediate level because the number of parts is small enough to group all of them in a single node.

Candidate parts: Candidates of atomic parts are represented with the position of the joint and its orientation vector. We consider all possible positions and orientations every 15 degrees. Since this label space is too large to apply our inference algorithm, for the first level we select the 300 candidates with lower unary potential. For the top level of the hierarchy, we select 500 candidates of hyperlabels with the branch-and-bound strategy. We set the prior $P(\ell)$ of Eq. (9) equal to 0 when upper arms and forearms have their shared boundaries too far away, or arms are too distant from the center of the torso. Otherwise, the prior was set to 1.

Unary potentials: We used the unary potentials of [25][3] for the first level, which are classification scores obtained with GentleBoost and HOG as descriptor. For the top level, we merged two unary potentials, the global shape model and the color correlation between parts, by learning the weighting factor. The global shape is modeled using the endpoint's coordinates of the segments of all parts (two pairs of coordinates per part, one at each extreme). We used the number of PCA eigenmodes that preserve 90% of the energy. For the color correlation potential we learned the relations with a linear SVM.

Consistency Potentials within Levels: In the proposed hierarchy, these potentials are only at the first level. We defined all of them as pairwise shape potentials between parts close to each other. Since none of them involves more than two atomic parts, we used the same Gaussian model as presented in [25].

[2]Available at http://www.robots.ox.ac.uk/∼vgg/data/stickmen.

[3]Available at http://vision.grasp.upenn.edu/cgi-bin/index.php?n=Video Learning.PSBaselineCode

	Torso	U. arm	Forearm	Head	Total
HCRF w/ shape+color model	99.7	91.3	63.9	95.5	**84.3**
HCRF w/ shape model	99.2	91.0	63.5	95.2	**83.9**
HCRF w/ color model	99.1	90.9	63.7	94.6	**83.8**
fully connected pairwise CRF	99.6	90.3	59.2	94.1	**82.1**
CRF (HCRF w/o top level)	99.7	88.7	55.8	93.0	**80.3**

TABLE I
Results on Buffy dataset.

Learning HCRF parameters: We learned the same parameters as with the facial feature selection experiment, *i.e.* the weighting of each potential and the Gaussian model of each pairwise shape potential. We also learned them with a Gibbs sampler of the parameter space.

B. Experiments

We use episodes 3 and 4 for training (472 images), and episodes 2, 5 and 6 for testing (276 images), as established for comparison with other methods. The evaluation criteria is the Percentage of Correctly detected Parts (PCP). A part is correctly localized if the endpoints of its segment lie within 50% of the ground-truth segment length (evaluation code supplied with the dataset).

Evaluation of the HCRF: In Table I we summarize the results of our HCRF and the CRF-based methods. Results show that we obtain a 4% of improvement when adding the top node, and 2% compared to the fully connected pairwise CRF. We believe that this improvement is because our HCRF is able to learn global relations among parts. Figure 5 shows some results.

At the top node, the candidates of [29] would represent the detection of the whole upper-body of the person. Because images in the Buffy dataset are already composed by the cropped upper-body, this method is unable to improve the results of pairwise CRF using our two-level hierarchy.

Comparison with state-of-the-art: The best results in the Buffy dataset are reported by [26] with a PCP of 85.5%. This method is based on a coarse-to-fine cascade of parts-based models. Each stage in the cascade prunes the pose space so that computationally expensive cues are computed at the final stages. It is important to notice that there is no theoretical reason why this strategy could not also be used in our generic framework.

VIII. Conclusions

We presented a novel HCRF for parts-based models fitting that uses product of label spaces to encode multiple atomic parts. Unlike pairwise CRFs, which only take into account local appearance and pairwise shape relations, our HCRF incorporates relations among sets of parts. This is specially important because it enables us to capture complex patterns such as the global shape structure or color correlations among parts. Experiments show that our HCRF obtains state-of-the-art results on facial feature detection and human pose estimation on two challenging datasets.

Appendix

Using the Bayes rule, the posterior becomes

$$P(\boldsymbol{\ell} = \mathbf{x}_i^* | \mathbf{O}) \propto P(\boldsymbol{\ell} = \mathbf{x}_i^*) P(\mathbf{O} | \boldsymbol{\ell} = \mathbf{x}_i^*). \quad (11)$$

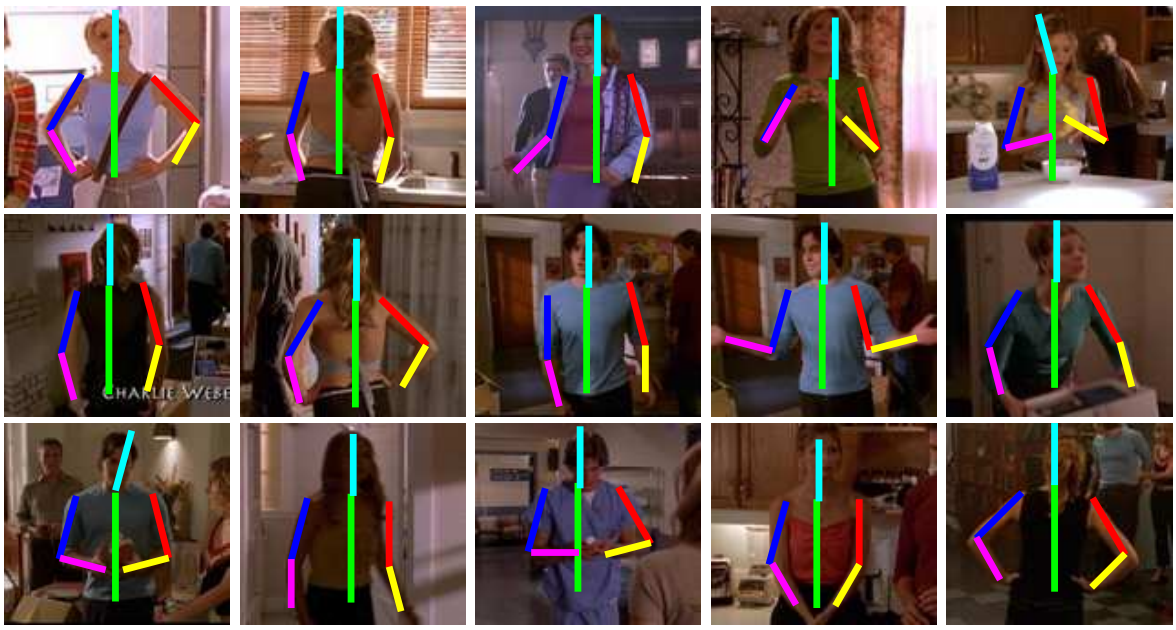

Fig. 5. Qualitative results on Buffy dataset.

The computation of the likelihood $P(\mathbf{O}|\boldsymbol{\ell} = \mathbf{x}_i^*)$ is hazardous due to the dimensionality of $\mathbf{O}$. We propose the following approximation:

$$P(\mathbf{O}|\boldsymbol{\ell} = \mathbf{x}_i^*) \approx P(\mathbf{O}_i|\boldsymbol{\ell} = \mathbf{x}_i^*) = \tag{12}$$

$$\prod_{p<s_i} P(\mathbf{O}_i^p|\ell^p = x_i^{p*}) \propto \tag{13}$$

$$\prod_{p<s_i} P(\ell^p = x_i^{p*}|\mathbf{O}_i^p), \tag{14}$$

where in Eq. (12) we discard the observations not relative to node $i \in \mathcal{V}$, in Eq. (13) we assume conditional independence between the observations done for each part $\mathbf{O}_i^p$, and finally, we apply again Bayes rule and assume equiprobability for all $\mathbf{O}_i^p$. Since inference will provide us a solution close to the optimal, we can easily compute the prior $P(\boldsymbol{\ell} = \mathbf{x}_i^*)$ as $P(\boldsymbol{\ell})$, and the factors $P(\ell^p = x_i^{p*}|\mathbf{O}_i^p)$ of the likelihood as $P(X_i^p = \ell^p|\mathbf{O}_i^p)$.

REFERENCES

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

[2] T. S. Caetano, T. Caelli, D. Schuurmans, and D. Barone. Graphical models and point pattern matching. *PAMI*, 2006.

[3] T. Cootes, Taylor, and C.J. Active shape models - 'smart snakes'. In *BMVC*, 1992.

[4] D. Cristinacce and T. Cootes. Boosted regression active shape models. In *BMVC*, 2007.

[5] F. de la Torre and M. Nguyen. Parameterized kernel principal component analysis: Theory and applications to supervised and unsupervised image alignment. In *Computer Vision and Pattern Recognition*, 2008.

[6] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *BMVC*, 2009.

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.

[8] V. Ferrari, M. M. Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

[9] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Pose search: retrieving people using their pose. In *CVPR*, 2009.

[10] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *TC*, 1973.

[11] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *IJCV*, 2000.

[12] J. M. Gonfaus, X. Boix, J. van de Weijer, A. Bagdanov, J. Serrat, and J. Gonzalez. Harmony potentials for joint classification and segmentation. In *CVPR*, 2010.

[13] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker. The CMU multi-pose, illumination, and expression (multi-PIE) face database. Technical report, Carnegie Mellon University Robotics Institute.TR-07-08, 2007.

[14] L. Gu, E. Xing, and T. Kanade. Learning GMRF structures for spatial priors. In *CVPR*, 2007.

[15] J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. *Unpublished*, 1971.

[16] A. Ihler and D. McAllester. Particle belief propagation. In *AISTATS*, 2009.

[17] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IT*, 2001.

[18] X. Lan and D. P. Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. In *ICCV*, 2005.

[19] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008.

[20] J. J. McAuley, T. S. Caetano, and M. S. Barbosa. Graph rigidity, cyclic belief propagation, and point pattern matching. *PAMI*, 2008.

[21] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006.

[22] X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV*, 2005.

[23] G. Roig, X. Boix, and F. De la Torre. Optimal feature selection for subspace image matching. In *WICCV*, 2009.

[24] A. P. S. Romdhani, S. Gong. A multi-view nonlinear active shape model using kernel PCA. In *ICCV*, 1999.

[25] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *CVPR*, 2010.

[26] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010.

[27] J. M. Saragih, S. Lucey, and J. F. Cohn. Face alignment through subspace constrained mean-shifts. In *ICCV*, 2009.

[28] E. B. Sudderth, A. T. Ihler, E. T. Ihler, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. In *CVPR*, 2002.

[29] L. Zhu, Y. Chen, and A. Yuille. Learning a hierarchical deformable template for rapid deformable object parsing. *PAMI*, 2010.