# Composite Templates for Cloth Modeling and Sketching

Hong Chen, Zi Jian Xu, Zi Qiang Liu and Song Chun Zhu
Departments of Statistics and Computer Science
University of California, Los Angeles
{hchen,zjxu,zqliu,sczhu}@stat.ucla.edu

## Abstract

*Cloth modeling and recognition is an important and challenging problem in both vision and graphics tasks, such as dressed human recognition and tracking, human sketch and portrait. In this paper, we present a context sensitive grammar in an And-Or graph representation which will produce a large set of* composite graphical templates *to account for the wide variabilities of cloth configurations, such as T-shirts, jackets, etc. In a supervised learning phase, we ask an artist to draw sketches on a set of dressed people, and we decompose the sketches into categories of cloth and body components: collars, shoulders, cuff, hands, pants, shoes etc. Each component has a number of distinct sub-templates (sub-graphs). These sub-templates serve as leaf-nodes in a big And-Or graph where an And-node represents a decomposition of the graph into sub-configurations with Markov relations for context and constraints (soft or hard), and an Or-node is a switch for choosing one out of a set of alternative And-nodes (sub-configurations) – similar to a node in stochastic context free grammar (SCFG). This representation integrates the SCFG for structural variability and the Markov (graphical) model for context. An algorithm which integrates the bottom-up proposals and the top-down information is proposed to infer the composite cloth template from the image.*

## 1.Introduction

Modeling human clothes is an important and challenging problem in many vision tasks that involve recognition of dressed people in natural environments, such as detecting[4, 11], tracking[5], surveillance, HCI, identification, human sketch and portrait for graphics rendering[2] etc. Despite intensive studies on the problems of "look at people" in the past decade, there is, to our best knowledge, no good model dedicated to realistic cloth modeling and recognition. The closest work that we can find are silhouette, contour, blob and region representations[6, 4]. In graphics, physical based models[8] are mostly used to create realistic visual effects of drapery and animation. Such models involve a large amount of polygons (mesh) and are very expensive. They do not account for the wide variability of cloth designs and are less relevant for the vision tasks mentioned above.

There are three major challenges in cloth representations.

*1. Geometric deformations*: clothes do not have static form and are very flexible.

*2. Photometric variabilities*: large variety of colors, shading effects, and textures.

*3. Topological configurations*: a combinatorial number of cloth designs – T-shirts, jackets, pockets, zips, suits, sweaters, coats where cloth components may be reconfigured and combined to yield new styles.

In this paper, we present a parsimonious yet expressive representation for clothes, motivated by the success of two sketch based methods. (i) Artists can draw concise sketches of clothes and human body that capture the most essential perceptual information[7]. (ii) The recent primal sketch model[3] – an attributed graph representation with a dictionary of image primitives aligned through landmarks – can reconstruct generic images with almost no perceptual distortions. We adopt a sketch graph representation like the primal sketch. Each stroke (long curves) of the sketch may correspond to (a) folds of clothes, (b) sewing lines, (c) occluding boundaries, and (d) shape outlines.

The geometric deformations of clothes are accounted for by the flexibility of the sketch graphs, and the photometric variabilities are accounted for by the rich image primitives whose parameters control the photometric variations. In the rest of the paper, we focus mostly on the study of topological configurations – a central theme in this paper, which has not been studied in the vision literature.

In a supervised learning phase, we collect a set of human images and sketches drawn by an artist. An example is shown in Fig 2.a-b. We remove one layer of the strokes that corresponds to shading folds and textures (Fig. 2.c). The remaining graph (Fig.2.d) is decomposed into a number of subgraphs for *cloth components* (Fig. 2.e). All subgraphs across the dataset, together with their neighborhood, are grouped into categories for collars, shoulders, cuff, hands,

pants, shoes and each has a number of possible structures (See Fig. 4). These subgraphs have "bonds" that tell them who to link with to compose bigger structures (see Fig. 6), and they are combined using the context information to form a wide variety of cloth *configurations*. For example, Fig. 5 shows three novel upper cloth configurations using some sub-templates in Fig. 4.

A crucial technique problem is: how can these components be composed into valid clothes? What are the rigorous mathematical models to govern the computation?

To account for the topological configurations, we build an And-Or graph, which is widely used in AI search [9], as an overall representation of clothes. An example is illustrated in Fig. 1. The And-Or graph is also used in the previous work [16] to present an attributed grammar in a generic rectangle parsing problem. The And-Or graph for clothes are quite different and more complicated. Each terminal (leaf) node (squares $1-11$) represents a component or sub-templates. Different sub-templates in the same category are represented by distinct leaves. The non-terminal nodes are divided into And-nodes whose children must be chosen jointly and Or-nodes of which only one child can be selected to express the alternative components. Intuitively, an And-node expands the configuration and an Or-node is a switch between alternative sub-configurations. The And-Or graph should be distinguished from a tree because the graph has horizontal edges (dashed) to specify the spatial relations and constraints among the components. A specific cloth configuration, say a jacket, corresponds to a subset of the And-Or graph (see the dark nodes and arrows). Thus one And-Or graph is like a "*mother template*" which produces a set of valid cloth configurations –"*composite templates*". A composite template is made of a set of leaf nodes (sub-templates). For example, leaf nodes $(1, 6, 8, 10)$ in Fig. 1 form a composite template. The spatial relations between the chosen leaf nodes are inherited from the And-Or graph. For example, the relation between nodes 1 and 6 is inherited from nodes B and C in Fig. 1, and the relation between nodes 6 and 8 is inherited from nodes N and O. These relations help to link the subgraphs (components) together to form a valid representation. In fact, the And-Or graph embodies a *context-sensitive grammar* which can generate a rich set of composite templates to account for the variabilities of clothes. It is a novel representational scheme that has not been studied yet in the vision literature. We will define a probability model on the context sensitive grammar in later section.

In a computing and recognition phase, we first activate some sub-templates in a bottom-up step. For example, we can detect the face and skin color to locate the coarse position of some components, which help to predict the positions of other components by context. Then a top-down step is activated to match some sub-templates in various

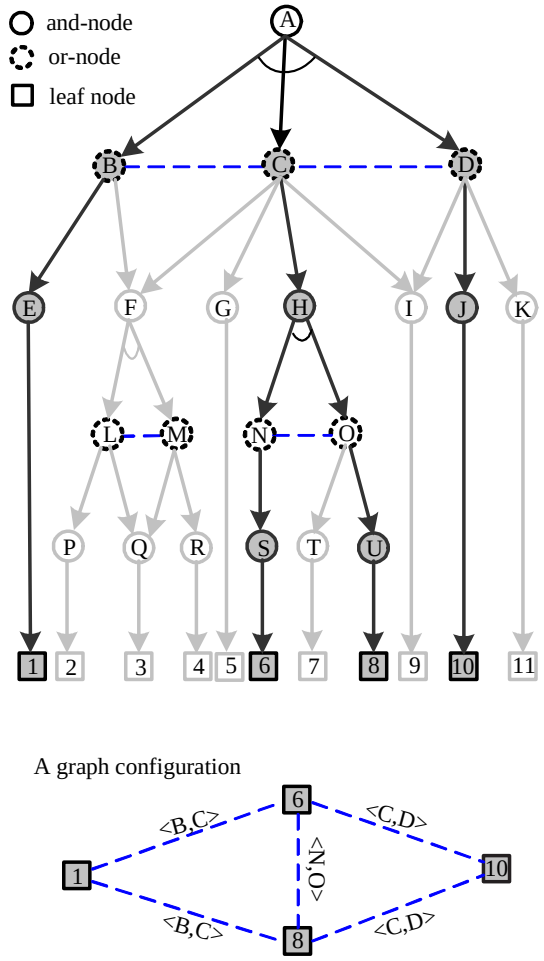
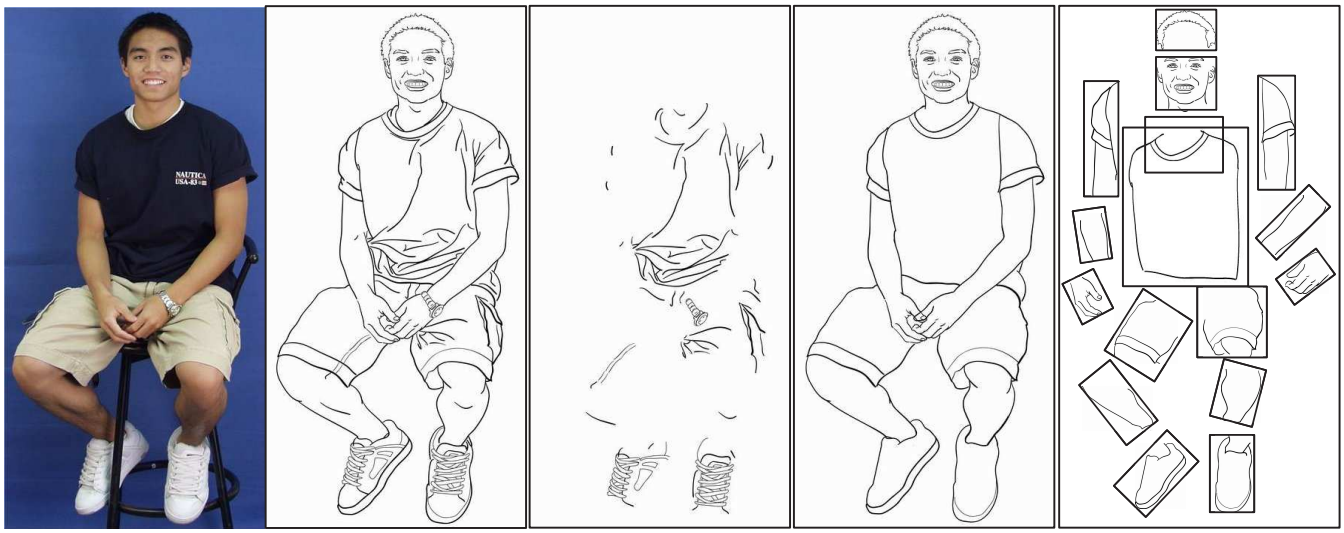



Figure 1. An illustration of the And-Or graph representation. The dark arrows and shaded nodes represent a composition of 4 leaf nodes (1,6,8,10) each being a *sub-template*. This generates a *composite graphical template* (at the bottom) representing the specific cloth *configuration* with the spatial relations (context) inherited from the And-Or graph.

categories to the edge maps. The matched graphs are deformed through diffusion to the exact boundaries in the image. Overall computation follows the Bayesian inference.

The context sensitive grammar and composite graphical template is a general modeling framework. We used cloth modeling and recognition as an example. This framework is applicable to many other classes of objects, especially classes where object instances have wide variabilities in their configurations, such as cars/van/truck, buildings, furniture and scene settings.

The paper is organized as follows. We first present the And-Or graph theory in Section 2 and a probabilistic context sensitive grammar model in Section 3 to set the theoretical foundation. Then we show the cloth model in Section 4 as an example, and briefly discuss the inference algorithm in 5. Some results are presented in Section 6 which is followed by a discussion on future work in Section 7.



(a). input image (b). artist sketch (c). folds/textures (d). structures (e). decomposition

Figure 2. An example of training the model. (a) Cloth image, (b) artist's sketch, (c) a layer of sketches for shading folds and texture (e.g. shoe lace, text on shirt etc), (d) remaining sketch graph for structures and sewing lines, (e) decomposed from (d) to sub-templates for components.

## 2.Composite graphical templates

Graphical templates are widely used in computer vision to model objects which have geometric deformation and photometric variabilities. A lot of object classes, such as clothes, have different topological configurations which cannot be modelled by a single graphical template. We propose to use a composite template to accommodate the topological variabilities, and we use the And-or graph as the overall representation of these composite graphical templates. The And-Or graph is a 5-tuple explained below.

$$\mathcal{G}_{\text{and}-\text{or}} =< N = U \cup V, T, \Sigma, \mathcal{R}, \mathcal{A} > . \quad (1)$$

1. *Non-terminal nodes* $N$ includes a set of And-nodes $U = \{u_1, ..., u_{m(U)})\}$ and a set of Or-nodes $V = \{v_1, ..., v_{m(V)}\}$. An And-node $u \in U$ represents a graph template which is composed of a set of sub-templates with certain relations $r_1, ..., r_k \in \mathcal{R}$ shown by the dashed horizontal lines in Fig. 1.

An Or-node $v \in V$ is a switch pointing to a number ($\geq 1$) of alternative sub-configurations. We define a switch variable $\omega(v)$ for $v \in V$, that takes an integer value as an index to the chosen node. $\omega(v) = \emptyset$ if $v$ is not used in the final chosen configuration.

2. *Terminal nodes* $T = \{t_1, ..., t_m(T)\}$ is a set of sub-templates $g_i$ representing object components for clothes, such as collars, pockets, hands, etc, as shown in Fig.3.

Each $g_i$ is a sketch graph where the sketch contours (strokes) are divided into many short segments and are connected by junctions. Both short segments and junctions are represented as vertices in $g_i$. Endpoints are 1-degree vertices, segments in the middle of contours are 2-degree vertices, and junctions have 3-4 degrees. Each vertex corresponds to an image primitive[3], such as step edge, bar, ridge, etc. We denote by $\mathbf{x}$ a vertex in $g_i$, $f$ an edge in $g_i$, and $\Lambda_i$ the image domain covered by $g_i$,

$$g_i = (\{\mathbf{x}_{i1}, ..., \mathbf{x}_{ik(i)}\}, \{f_{mn} =< \mathbf{x}_{im}, \mathbf{x}_{in} >\}, \Lambda_i). \quad (2)$$

3. *Configurations* $\Sigma$ is a finite set of valid composite templates,

$$\Sigma = \{G_j = (g_{j,1}, ..., g_{j,m(j)}) : j = 1, 2, ..., M\}. \quad (3)$$

Each graph $G \in \Sigma$ is a specific configuration for the object, such as a suit, a jacket or a T-shirt. For example $G = (1, 6, 8, 10)$ is a configuration in Fig.1 (bottom). The spatial relations/constraints between these leaves are inherited from their parents in the graphs. For example, the relation between node 6 and node 8 are inherited from nodes N and O. The And-Or graph in Fig.1 contains a combinatorial number of valid configurations, e.g.

$$\Sigma = \{(1, 6, 8, 10), (1, 5, 11), (2, 4, 6, 7, 9), ...\} \quad (4)$$

The number of nodes may vary as well as the graphical structures. Fig. 5 shows three novel upper cloth configurations generated by the And-Or graph of clothes.

4. $\mathcal{R}$ is a set of relations between Or-nodes or sub-graphs.

$$\mathcal{R} = \{r_{ij} =< v_i, v_j >: v_i, v_j \in V\}. \quad (5)$$

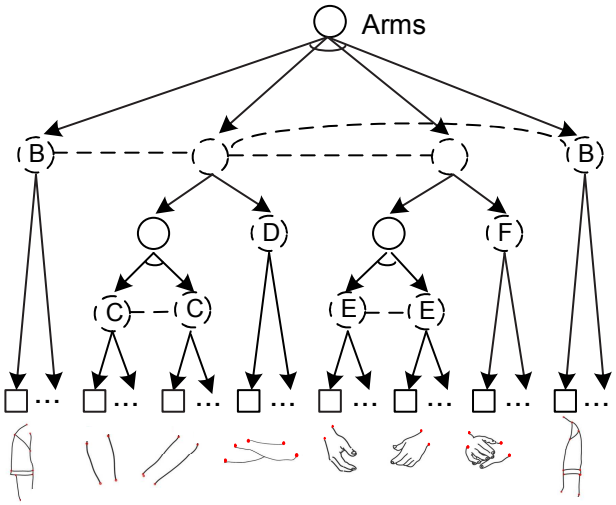



Figure 3. The And-Or graph for arms.

These relations become the pair-cliques in the composite graphical template. When a node $v_i$ is split later, the relation $r_{ij}$ will be split as well. For example, in Fig. 1 node C is split into two leaf nodes 6 and 8, then the relation $< B, C >$ is split into two subsets between 1-6 and 1-8,

$$< B, C >=< 1, 6 > \cup < 1, 8 > . \tag{6}$$

5. *Attributes* $\mathcal{A}$ is a set of photometric and geometric transforms applied to the sub-templates $g_i$ on the leaf nodes $t_i \in T, i = 1, ..., m(T)$.

$$\mathcal{A} = \{(\mathcal{A}_i^{\text{pho}}, \mathcal{A}_i^{\text{geo}}) : i = 1, 2, ..., m(T)\}. \tag{7}$$

The photometric attributes $\mathcal{A}_i^{\text{pho}}$ is a vector – the type of intensity/color profiles and their contrasts – for the primitives of the vertices $\mathbf{x}_{i1}, ..., \mathbf{x}_{ik(i)}$ in sub-graph $g_i$. The geometric transforms include affine transform $A_i = (Tx, Ty, s, \theta)$ for the whole subgraph $g_i$ and warping (deformation or displacement) $(\xi_{ij}, \eta_{ij})$ for each vertex.

$$\mathcal{A}_i^{\text{geo}} = (A_i, (\xi_{i1}, \eta_{i1}), ..., (\xi_{in(i)}, \eta_{ik(i)})). \tag{8}$$

The photometric and geometrical transforms, together with the topological variations construct a very rich set of composite (attributed) templates, and the graph matching algorithm in later section will be defined on the photometric, geometric, and topological distances.

## 3.Context sensitive grammar models

Stochastic context free grammars (SCFG) were introduced to vision in (Fu, 1981)[1], and are equivalent to the Markov tree model. Such grammar models suffer a major problem that they cannot pass global and context information among nodes. For example, they cannot put a constraint that a face consists of 2 (not 3 or 4) nearly symmetric eyes. The context information is best described by graphic templates (Yuille 1991)[13] and Markov random fields (MRF) on graphs. But a fixed template cannot account for the structural variabilities. The And-Or graphs integrate both representations.

The And-Or graph presented in the previous section is a combination of a Markov tree and Markov random fields. Each And-node is a graphic (MRF) template model, and each Or-node is a switch node in the Markov tree (SCFG). As a matter of fact, each And-Or graph $\mathcal{G} =< U \cup V, T, \Sigma, \mathcal{R}, \mathcal{A} >$ is a *context sensitive grammar*[10] where $T$ is the vocabulary, $U$ and $V$ are two types of production rules ($U$ for graph expansion with relation $R$ and $V$ for graph switches), $\Sigma$ is a language (sentences) generated by these rules, and $\mathcal{A}$ is the photometric and geometric attributes associated with the vocabulary. It can generate a large set of configurations (sentences) like the SCFG, and at the same time each configuration $G \in \Sigma$ carries the context and Markov constraints (soft or hard) in a graphic model.

The context sensitive grammar will be useful for many vision tasks. Cloth modeling in the this paper is just one example. In the following, we define a probabilistic model $p$ on top of the composite graphical templates $\Sigma$ for modeling and inference.

Let $G \in \Sigma(\mathcal{G})$ be a composite template. It has the following constituents.

1. $V(G) = \{v_1, ..., v_m\}$ is a set of Or-nodes (switches) that are used in configuring $G$. For instance, $V(G) = \{B, C, D, N, O\}$ in Fig.1 for the configuration $G =$(1,6,8,10).

2. $T(G) = \{t_1, ..., t_n\}$ is the leaf nodes in configuration $G$, such as (1,6,8,10) in Fig.1 . Each is a subgraph $g_i(t_i), i = 1, 2, ..., n$.

3. $\mathcal{R}(G) = \{r_{i,j} =< g_i, g_j >\}$ is the set of relations defined on terminal sub-templates inherited from the And-Or graph $\mathcal{G}$. Each is a pair-clique and $g_i$ and $g_j$ are both terminal node in the And-Or graph.

4. $S(G)$ is the photometric and geometric transforms applied to the subgraphs $g_i, i = 1, 2, .., n$.

The probability for $G$ is

$$p(G; \mathcal{G}) = \frac{1}{Z(\mathcal{G})} \exp\{-E(G)\} \tag{9}$$

$$E(G) = \sum_{v \in V} E_v(\omega(v)) + \sum_{t_i \in T(G)} E(g_i) + \sum_{r_{i,j} \in \mathcal{R}(G)} E(g(v_i), g(v_j)). \tag{10}$$

The first term in the energy $E$ is the same as the SCFG. It assigns different weights to the switch variables $\omega(v)$ at the or-nodes $v$, and each accounts for how frequently an And-node appears. Removing the 2nd and 3rd terms, this reduces to a SCFG.

The second and third terms are typical singleton and pair-clique energy defined on the graph $G$ *after the switch variables are decided*. The second term is the prior model

of the geometric and photometric transformations applied to the sub-template. The third term models the compatibility constraint, such as the spatial constraint between sub-templates. We will give the detailed energy of cloth modelling in following section.

In the above model, the partition function is related to the and-Or graph $\mathcal{G}$ and is common to all graph configurations in $\Sigma(\mathcal{G})$.

$$Z(\mathcal{G}) = \sum_{G \in \Sigma(\mathcal{G})} \exp\{-E(G)\}. \tag{11}$$

Because of a common $Z$, we no longer need to worry about computing the partition function or ratio when we switch between different configurations $G \leftrightarrow G'$ in the inference phase.

## 4.Composite models of clothes

We take 50 training images of college students sitting in a high chair with good light conditions and uniform background to exclude occlusions and bad illuminations. Gesture, illuminations and background clutter are other difficult problems that are beyond the scope of this paper. This paper is focused on modeling the variability of cloth configurations. An artist is paid (in hourly rate) to draw the sketches in Adobe illustrator. She is asked to make the sketches as consistent as possible across the training images.

As Fig. 2 shows, we first manually separate a layer of sketches corresponding to shading folds and textures (e.g. shoe lace, text printed on T-shirt). Then we decompose the remaining structures (Fig. 2.d) into a number of sub-templates: hair, face, collar, shoulder, upper and lower arms, cuff, hands, pants, shoes, and pocket. Fig. 4 shows some examples for each category.

With these categories, we construct an And-Or graph to account for the variability of configurations. For an example, we shown the And-Or graph of arms in Fig. 3. Intuitively, the And-Or graph is like a "mother template" which can produce a large set of configurations, three of which are shown in Fig. 5.

For each terminal node, we assume uniform probability for choosing the primitives and intensity profile and put a Thin Plate Spline (TPS) model to regularize the warping (deformation controlled by the warping on the vertices) $(\xi, \eta)$ of subgraphs $g_i$ on domain $\Lambda_i$. Therefore, the singleton $E(g_i)$ in Eqn.10 is defined by,

$$E(g_i) = \int\int_{\Lambda_i} \xi_{xx}^2 + \xi_{xy}^2 + \xi_{yy}^2 + \eta_{xx}^2 + \eta_{xy}^2 + \eta_{yy}^2 dxdy. \tag{12}$$

We define a set of "bonds" for each sub-template. They are used to link the corresponding "bonds" in neighbor parts, such as the torso and upper arm or the upper arm and the forearm. Let's denote the set of all connection of "bonds" between sub-template $g_i$ and $g_j$ as $B_{i,j}$. Then, the

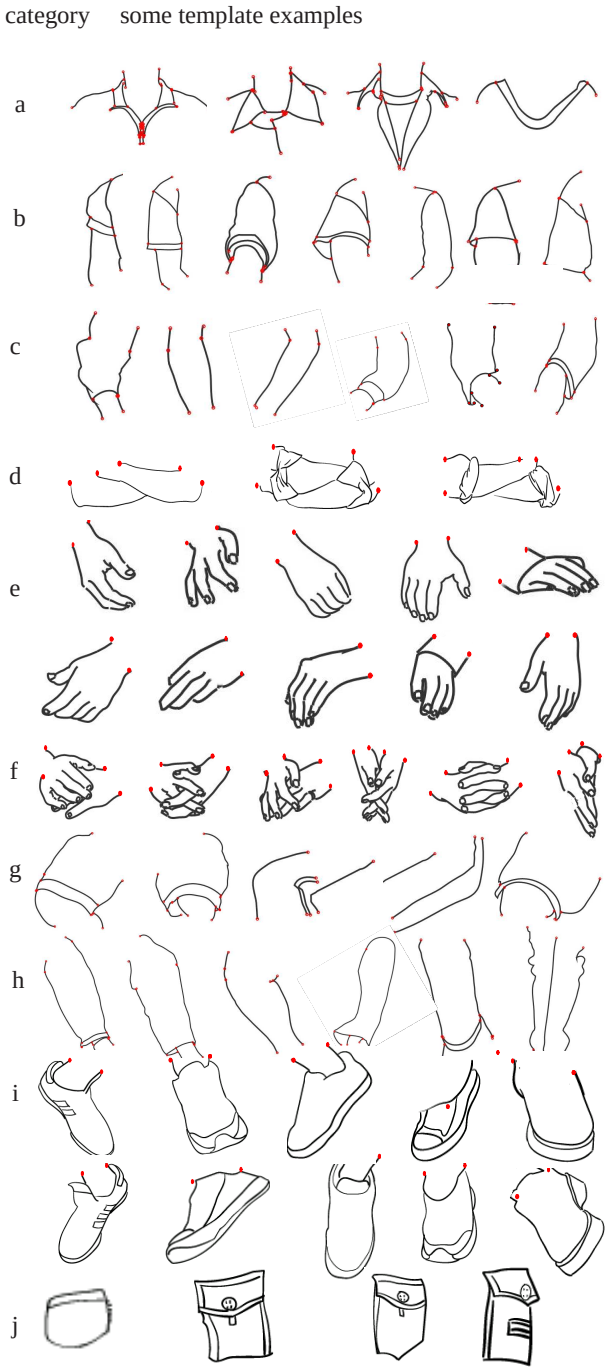



Figure 4. The categories for cloth components and each category consists of a set of sub-templates used as leaf nodes in the And-Or graph.

pair-clique term is defined as

$$E(g_i, g_j) = d(A_i, A_j) + \sum_{<\beta_{ik}, \beta_{jl}> \in B_{i,j}} d(\mathbf{x}(\beta_{ik}), \mathbf{x}(\beta_{jl})).$$

, where the first term is used to enforce consistence of the

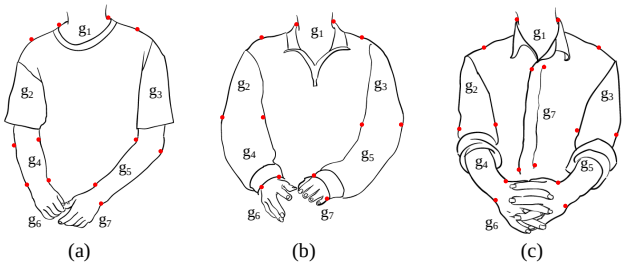



Figure 5. Three novel configurations composed of 6,5,7 sub-templates in the categories respectively. The bonds are shown by the red dots.

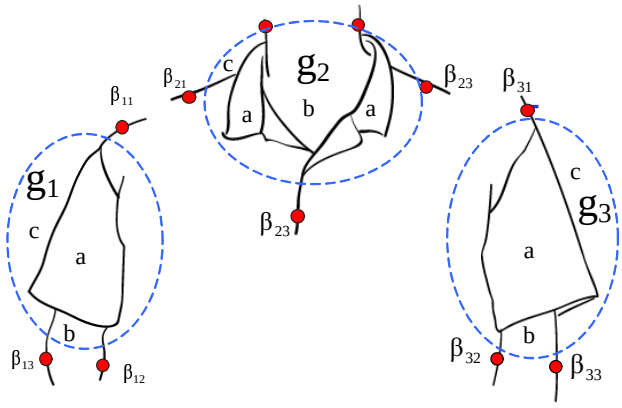



Figure 6. Three subgraphs $g_1, g_2, g_3$ (inside the ellipses) and their environments $\text{env}(g_1), \text{env}(g_2), \text{env}(g_3)$ (outside the ellipses). The bonds are shown by the red dots. $g_1$ has three bonds $\beta_{11}, \beta_{12}, \beta_{13}$. The relation between $g_1$ and $g_2$ are specified in $(\beta_{11}, \beta_{21})$.

affine transforms of the two sub-templates.

$$d(A_i, A_j) = \lambda_s(\log(s_i/s_j) - \alpha_s)^2 + \lambda_\theta \sin((\theta_i - \theta_j - \alpha_\theta)/2)^2.$$

,where $s, \theta$ are the affine transformation parameters for each part, and $\lambda_s, \alpha_s, \lambda_\theta, \alpha_\theta$ are learned parameters.

The second term is used to enforce the corresponding "bonds" between neighboring parts, which are connected with each other.

The likelihoods are models from sub-template $g_i$ with domain $\Lambda_i$ to the image patch $\mathbf{I}_{\Lambda_i}$. The likelihood model is

$$p(\mathbf{I}_\Lambda | G; \Delta) = \prod_{g_i \in T(G)} p(\mathbf{I}_{\Lambda_i} | g_i; \Delta_i), \qquad (13)$$

where $\Delta = \{\Delta_1, ..., \Delta_N\}$ is the set of all image primitives[3] associated with the templates. Within each domain $\Lambda_i$, some pixels are covered by the image primitives (for vertices) and the remaining part are filled in by MRF models as in the primal sketch[3]. People who are not familiar with the primal sketch should think of it like inpainting. The background, which is not covered by our templates, is modeled as uniform Gaussian noise. For each template, we will record the following attributes: (a) the type of boundary, (b) ownership (which side owns an occluding boundary), (c) the intensity profiles and types of image primitives, (d) region model (illustrated by a-b-c in Fig. 6). Since the likelihood model for graphical templates are straightforward, we choose not to unfold its details due to space limits.

The parameters in the above models are learned independently from the set of training examples.

## 5.Bottom-up and top-down inference

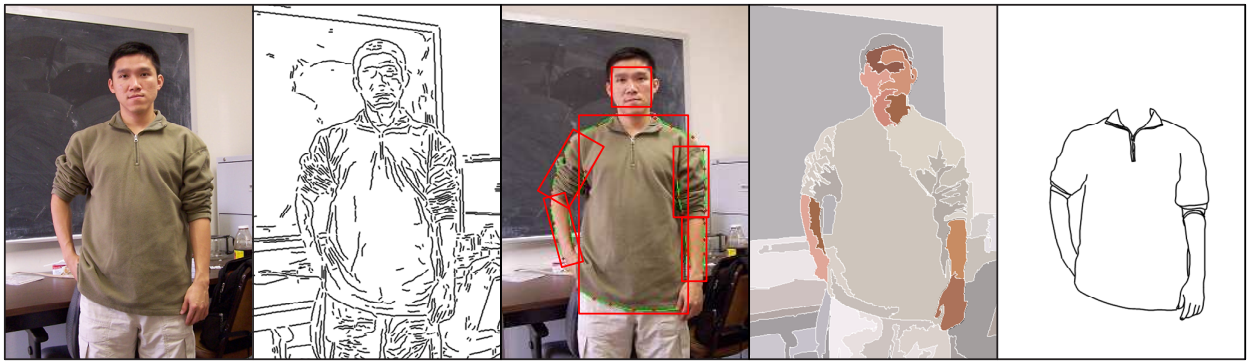
Given the prior model with all the descriptions $\mathcal{G}$ and the generative likelihood model, the objective of the inference algorithm is to compute the best graph configuration $G \in \Sigma(\mathcal{G})$ for a test image $\mathbf{I}$,

$$G^* = \arg \max_{G \in \Sigma(\mathcal{G})} p(\mathbf{I}_\Lambda | G; \Delta) p(G). \qquad (14)$$

The inference is done in a spirit similar to the DDMCMC image parsing[12] which combine the bottom-up and top-down computation.

Bottom-up data driven proposals (discriminative tests) are designed for all nodes in the And-Or graph. Some nodes are more informative and thus can be detected more reliably, such as the face. Other nodes are less informative. For instance, the elbow is hard to infer when the arm is straight. It is also desirable to infer the nodes from coarse-to-fine. For example, the leaf nodes have many sub-templates which should be activated after their positions are located (predicted) from the parent nodes in the higher level of the And-Or graph. We carefully design the order of the bottom-up and top-down computation so that the more informative nodes are proposed and inferred earlier, and they in turn generate useful top-down/context information for computing the other less informative nodes.

Fig. 7 shows an example of our bottom-up and top-down inference. We do face detection as shown in Fig. 7.(c). To locate the torso, we use an ASM (Active Shape Model)[17] type of model. We define a common shape template for the key points of the torso as the red points in Fig. 7.(c). We apply PCA to the shape and model the prior of shape as a Gaussian distribution in the reduced dimension. With the position of the face, and the relative position model of torso and face, we randomly sample some shapes to use as the initialization of the ASM searching algorithm. Using the same method, we build shape models for upper arms and forearms. Note in Fig.3, there are two distinct configurations of the forearms. We use a mixture model for the two distinct configurations of forearms: separated or crossed. After locating these body parts with bottom-up methods, we obtain an overall proposal for body parts as shown in Fig. 7(c). The face rectangle is used as the initialization of the skin region,



(a) input image (b) edge map (c) face and parts (d) skin color (e) result

Figure 7. Running example. (a) input image, (b) Canny edge map, (c) bottom-up detection of face and body parts, (d) skin color detection by mean-shift clustering in color space, (e) results of top-down matching.

and combined with the results of mean shift clustering, we get the skin region as in Fig. 7.(d).

With the coarse positions of the human parts, we start the top-down graph matching process. Each non-terminal node at the bottom of the And-Or graph has a number of leaf nodes (see Fig.3) which correspond to a set of sub-templates (subgraphs). Each sub-template is a subgraph with photometric and geometric transformation parameters. Deformable Template matching is well studied, especially with good initialization. For each candidate sub-template, we match it to the image with the thin-plate spline prior (defined in the prior terms in eqn.(12)) with iterated closest point algorithm (*TPS-ICP*)[14]. The best matched sub-template is selected by comparing the posterior probability. Although the TPS-ICP algorithm is a locally optimal algorithm, the likelihood models with region information in the sub-templates are quite robust. An example of the fitting result is shown in Fig. 7.(e).

## 6.Experiments

We applied our algorithms to two sets of images. One set is in Fig.8, and another set is in Fig.9. For each testing image, as in Fig. 7, we first infer the body gesture at the parts layer, then infer the composite graphical template. As we can see, the whole hand is represented by one or two sub-templates. We are not computing the fingers individually which seems an impractical task. In case a hand in the test image has a different configuration from the sub-templates in our training set, we choose the closest match. It is interesting to observe that human vision is apparently not very sensitive to subtle differences.

As shown in Fig. 8 and Fig.9, these graphical sketches are quite nice for they are generated from the artist's templates. One can use such results in many applications, such as cartoon animation, human portraits, and video communication in narrow band devices.

Figure 8. Some recognition results of upper body with clothes. The input images are shown in the first row. The third row shows the composite graphical templates inferred from the images. For comparison, the results of a canny edge detector are shown in the middle row.

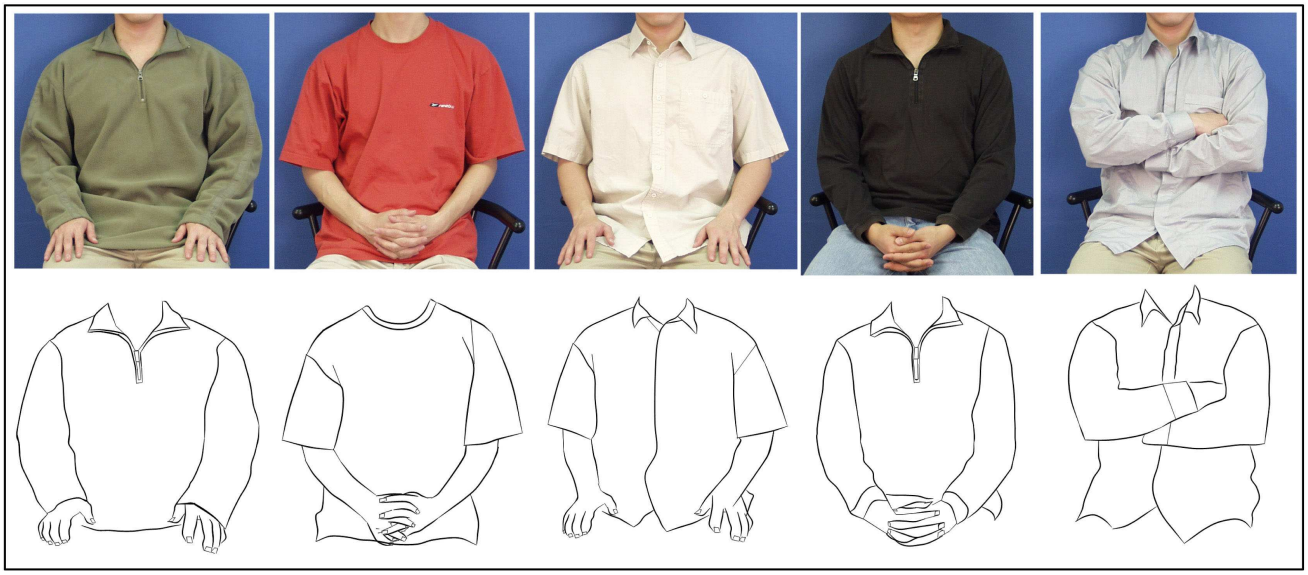


Figure 9. More recognition results of upper body with clothes. The results are the composite graphical templates inferred from the images.

## 7.Discussion

The main contribution of this paper is the And-Or graph representation as a rigorous matching model for generating a set of composite graphical templates in a principled way, in contrast of the graphical (MRF) models/template of fixed structure widely used in the literature. This representation embodies the context sensitive grammar model which was never used in the vision literature before, despite the fact that people have long desired such a model. We used cloth modeling and recognition as an example, however the framework is generally applicable to many other classes of objects, especially classes where object instances have wide variabilities in their configurations.

Due to space limit, we cannot unfold many details on graph matching and the likelihood models, which are, in our opinion, straightforward and not new in this paper. We refer to the literature for existing work on graph matching and primal sketch[3] for likelihood models.

## Acknowledgement

This work is supported by grants ONR N-0014-05-1-0543, NSF IIS-0413214 and Kodak Fellowship.

## References

[1] K.S. Fu, *Syntactic Pattern Recognition and Applications*, Prentice Hall, 1981.

[2] H. Chen and Z.Q. Liu, C. Rose, Y.Q. Xu, H.Y. Shum and D. Salesin, "Example-Based Composite Sketching of Human Portraits", *Proc. 3rd Int'l Symp. NPAR*, 95-153, 2004.

[3] C.E. Guo, S.C. Zhu, and Y.N. Wu, "A mathematical theory for primal sketch and sketchability," *ICCV*, 2003.

[4] S. Ioffe and D. A. Forsyth, "Probabilistic methods for finding people", *IJCV*,43(1):45 68, 2001.

[5] M. W. Lee and I. Cohen, "Human upper body pose estimation in static images", *ECCV*, 2004.

[6] L. Zhao, "Dressed Human Modeling, Detection, and Parts Localization", *Ph.D Dissertation*, CMU, July, 2001.

[7] K. Nicolaides, *The Natural Way to Draw*, Houghton Mifflin Co. Boston, 1941.

[8] H. N. Ng et al, "Computer Graphics Techniques for Modeling Cloth", *IEEE Computer Graphics & Applications* , V16, pp 28-42, 1996.

[9] J. Pearl, *Heuristics: Intelligent Search Strategies for Computer Problem Solving*, Addison-Wesley, 1984.

[10] J. Rekers and A. Schürr, "A parsing algorithm for context sensitive graph grammars", TR-95-05, Leiden Univ. 1995.

[11] R. Ronfard, C. Schmid and B. Triggs, "Learning to parse pictures of people", *ECCV*, 2002.

[12] Z.W. Tu et al. "Image parsing: unifying segmentation, detection, and recognition", *ICCV*, 2003.

[13] A.L. Yuille, D. Cohen, and P. Hallinan, "Feature extraction from faces using deformable templates", *IJCV*, vol.8, 99-111,1992.

[14] H. Chui and A. Rangarajan, "A new algorithm for non-rigid point matching", *CVPR*,2000.

[15] G. Hua, M. H. Yang, Y. Wu, "Learning to estimate human pose with data-driven belief propagation", *CVPR*, 2005.

[16] F. Han and S. C. Zhu , "Buttom-up/Top-down image parsing by attribute graph grammar", *ICCV*, 2005

[17] T.F.Cootes and C.J.Taylor, "Statistical models of appearance for computer vision", *Tech. Report*, Univ. of Manchester, U.K., 2000.