ImageSpirit: Verbal Guided Image Parsing

Ming-Ming Cheng, Shuai Zheng, WenYan Lin, Jonathan Warrell, Vibhav Vineet, Paul Sturgess, Nigel Crook
Oxford Brookes University
Niloy J. Mitra
University College London
and Philip H. S. Torr
University of Oxford

Humans describe images in terms of nouns and adjectives while algorithms operate on images represented as sets of pixels. Bridging this gap between how we would like to access images versus their typical representation is the goal of image parsing. In this paper we propose treating nouns as object labels and adjectives as visual attributes. This allows us to formulate the image parsing problem as one of jointly estimating per-pixel object and attribute labels from a set of training images. We propose an efficient (interactive time) solution to this problem. Using the extracted attribute labels as handles, our system empowers a user to verbally refine the results. This enables hands free parsing of an image into pixel-wise object/attribute labels that correspond to human semantics. Verbally selecting objects of interests enables a novel and natural interaction modality that can possibly be used to interact with new generation devices (e.g., smart phones and Google glasses). We demonstrate our system on a large number of real-world images with varying complexity and understand the tradeoffs compared to traditional mouse-based interactions using both a user study and large scale quantitative evaluation.

Categories and Subject Descriptors: I.3.6 [**Computer Graphics**]: Methodology and Techniques—*Interaction Techniques*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object Recognition*

General Terms: image parsing, nature language control

Additional Key Words and Phrases: object class segmentation, semantic attributes, multi-label CRF

1. INTRODUCTION

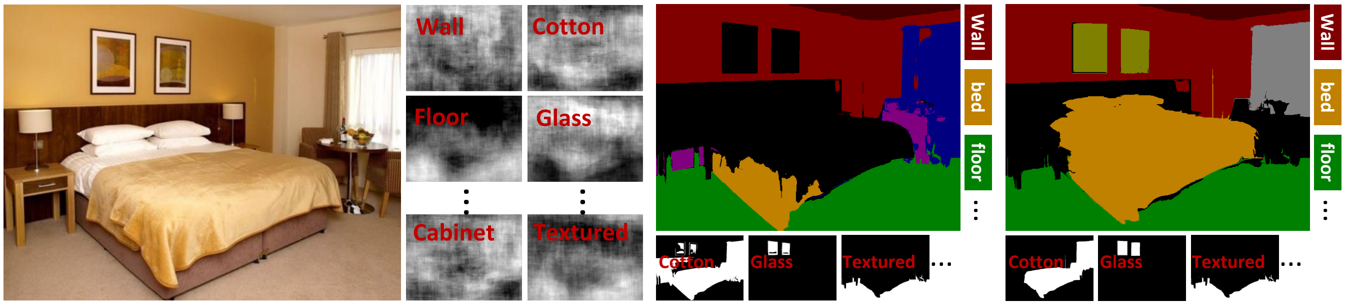
Humans perceive images in terms of language components of nouns (e.g., bed, cupboard, desk) and adjectives (e.g., textured, wooden). In contrast, pixels are nature representations for computers [Ferrari and Zisserman 2007]. Bridging this gap between our mental models and machine representations is the domain of image parsing. This consists of two key components: image segmentation and the assignment of verbal tags such as object names/attributes to the segments. This is a difficult problem. While to date, there exist large numbers of automated image parsing techniques [Ladicky et al. 2009; Shotton et al. 2009; Kulkarni et al. 2011; Tighe and Lazebnik 2011; Krähenbühl and Koltun 2011], the methods often require manual correction especially in real-world images. In this paper, we propose an efficient approach that allows users to produce high quality image parsing results by simply talking to the software. This enables hands free parsing of an image into pixel-wise object labels that are meaningful to both humans and computers. The output could directly be consumed by new generation of devices such as smart phones and Google glasses, which do not readily accommodate mouse interaction. Such an interaction modality not only enriches how we interaction with the images, but also provides important interaction ability for applications where non-touch manipulation is crucial [Hospital 2008] or hands are busy in other ways [Henderson 2008].

We face three technical challenges in developing verbal[1] guided scene parsing: (i) words are amorphous concepts that are difficult to translate into pixel level meaning; (ii) how to control the overall system using only verbal cues, and (iii) ensuring the system responds at interactive rates. To address the first problem, we treat nouns as objects and adjectives as attributes. Using training data, we obtain a pixel-wise hypothesis for each object and attribute, e.g., Figure 1(a). These are integrated through a novel, multi-label factorial conditional random field (CRF) that jointly estimates both object and attribute segmentation as seen in Figure 1(b). Not only does this joint segmentation provide verbal handles to the underlying image, the ability of object and attribute labels to reinforce each other results in a better overall solution when compared to prior object-only segmentation techniques [Ladicky et al. 2009; Krähenbühl and Koltun 2011]. Our second problem of verbal control is also naturally addressed by our joint object and attribute CRF. We use adjectives in the users command as automatic attribute predictions and the correlation between adjective/attributes and nouns/objects to mutually reinforce each other. This allows the user to intuitively incorporate high level understanding of the current image and quickly find discriminative visual attributes to improve scene parsing. This can be seen in Figure 1(c) where given verbal inputs, such as 'glass picture', our algorithm can re-weight the CRF's for both glass and picture to provide a good quality 'picture' segment boundary. Finally, we show our joint CRF formulation can be factorized for

Authors' addresses: cmm.thu@gmail.com.

DOI 10.1145/XXXXXXX.YYYYYYY
http://doi.acm.org/10.1145/XXXXXXX.YYYYYYY

[1] We use the term verbal as a short hand to indicate word-based, i.e., nouns, adjectives, and verbs. We make this distinction as we focus on semantic image parsing rather than speech recognition or natural language processing.



(a) Inputs: an image and learned weak hypothesis [Shotton et al. 2009] (b) Automatic scene parsing results (c) Nature language guided parsing

Fig. 1. Given a source image downloaded from the Internet, our system generates multiple weak object/attributes cues (a). Using a novel multi-label CRF, we generate per-pixel object and attribute labeling (b). Based on this output, additional verbal guidance: 'Refine the cotton bed in center-middle', 'Refine the white bed in center-middle', 'Refine the glass picture', 'Correct the wooden white cabinet in top-right to window' allows re-weighting of CRF terms to generate, at interactive rates, high quality scene parsing result (c). Best viewed in color.

efficient inference. This permits the use of efficient filtering based techniques to perform inference at interactive speeds.

For evaluation, we created a system for parsing indoor scenes. Training data was obtained from an augmented NYU v2 RGB image dataset that contains 1449 indoor images. By splitting this data into training and test sets, we performed empirical evaluation with respect to the state of the art object segmentation algorithms. We report a 6% improvement using our automated object/attribute segmentation. Beyond these numbers, our algorithm provides critical verbal handles for refinement and subsequent edits. This leads to a very large (30%) improvement if verbal interaction is added. These outputs also correspond more closely to human perception than traditional automatic object segmentation and are more easily integrated into subsequent applications. This hypothesis is validated by extensive evaluation results provided in the supplementary material. Outside of the database, we find that our system generalizes to other images of similar scene type. Thus our indoor scene parsing system can work on images downloaded from the Internet using for example, 'bedroom' as a search word.

While scene parsing is important in its own right, we believe our system can enable novel human-computer scene interactions. Specifically, by providing hands free selection mechanics to indicate objects of interest to the computer, we can largely replace the role traditionally filled by the mouse. This enables interesting image editing modalities such as speech guided image manipulation and can be integrated in smart phones and Google glasses, by making commands such as 'zoom in on the cupboard in the far right' meaningful to the computer.

In summary, our main contributions are:

(1) a new interaction modality that enables language command to guide image parsing;
(2) the development of a novel multi-label factorial CRF that can integrate cues from multiple sources at interactive rates; and
(3) a demonstration of the potential of this approach to make conventional mouse based tasks hands-free.

2. RELATED WORKS

Object class image segmentation and visual attributes. Assigning an object label to each image pixel, known as object class image segmentation or scene parsing, is one of computer vision's core problems. TextonBoost [Shotton et al. 2009], is a ground breaking work for solving this problem. It simultaneously achieves pixel level object class recognition and segmentation by jointly modeling patterns of texture and their spatial layout. Several refinements of this method have been proposed, including context information modeling [Rabinovich et al. 2007], joint factorial CRF [Ladicky et al. 2010], dealing with partial labeling [Verbeek and Triggs 2007], and efficient inference [Krähenbühl and Koltun 2011]. These methods deal only with noun like object labels, and not adjectives. Further they require each pixel to take only one label. Visual attributes [Ferrari and Zisserman 2007], which describe important semantic properties of objects, have been shown to be an important factor for improving object recognition [Farhadi et al. 2009; Wang and Mori 2010], scene attributes classification [Patterson and Hays 2012], and even modeling of unseen objects [Lampert et al. 2009]. So far the work on attributes has been limited to determining the attributes of an image region contained in a rectangular bounding box. Recently, Tighe and Lazebnik [2011] have addressed the problem of parsing image regions with multiple label sets. However, their inference formulation remain unaware of object boundaries and the obtained object labeling usually spreads the entire image. We would like to integrate both object and attribute segmentation. This is a very difficult problem as, in contrast to traditional segmentation in which only one label is predicted per pixel, there now might be zero, one, or a set of labels predicted for each pixel, e.g., a pixel might belong to wood, brown, cabinet, and shiny. Our model is defined on pixels with fully connected graph topology, which has been shown [Krähenbühl and Koltun 2011] to be able to produce fine detailed boundaries.

Interactive image labeling. Interactive image labeling is an active research field. This field has two distinct trends. The first involves having some user defined scribbles or bounding box which is used to assist the computer in cutting out the desired object from image [Mortensen and Barrett 1995; Liu et al. 2009; Li et al. 2004; Blake et al. 2004; Rother et al. 2004; Lempitsky et al. 2009]. Gaussian mixture models (GMM) are often employed to model the color distribution of foreground and background. Final results are achieved via Graph Cut [Boykov and Jolly 2001]. While widely used, these works do not extend naturally to verbal segmentation as the more direct scribbles cannot be replaced with vague verbal descriptions such as 'glass'. The second interaction type involves adding a human-in-the-loop [Branson et al. 2010; Wah et al. 2011].

Such works focus on recognition of image objects rather than segmentation. They resolve ambiguities by prompting the user for simple inputs. For example, if recognizing birds, they will seek user guidance through questions such as 'are the feather's red?' or 'circle the beak'. Our work can be considered a verbal guided human-in-the-loop semantic segmentation. However, our problem is more difficult than the usual human-in-the-loop problems because of the ambiguity of words (as opposed to binary answers to questions) and the requirement for fine pixel wise labeling (as opposed to categorization). This precludes usage of a simple tree structure for querying and motivates our more sophisticated, interactive CRF model to resolve the ambiguities.

Semantic-based region selection. Manipulation in the semantic space is a powerful tool and there are a number of approaches that treat this as an image-retrieval problem through some user annotation. An example is Photo Clip Art [Lalonde et al. 2007] which allows users to directly insert new semantic objects into existing images, by retrieving a suitable objects from image based object database. Chen *et al.*[2009] who further extended this work to sketch based image composition by automatically extracting and selecting suitable objects candidates from Internet images. Zhou *et al.*[2010] proposed to reshape human image regions by fitting an appropriate 3D human models. Zheng *et al.*[2012] partially recovered the 3D of man-made environments, enabling intuitive non-local editing. However, none of these methods, attempt interactive verbal guided image parsing which has the added difficulty of an image containing multiple objects and verbal commands being vague guidance cues.

Speech interface. Speech interfaces are deployed when mouse based interactions are infeasible or cumbersome. Although research on integrating speech interfaces into software started in the 1980s [Bolt 1980], it is only recently that such interfaces have been widely deployed, (e.g. Apple's Siri, PixelTone [2013] and Google voice search). However, most speech interface research is focused on natural language processing and to our knowledge there has been no prior work addressing image region selection through speech. The speech interface that most resembles our work is PixelTone [2013], which allows users to attach object labels to scribble based segments. These labels allow subsequent voice reference. Partly inspired by PixelTone, we have developed an entirely hands free parsing of an image into pixel-wise object/attribute labels that correspond to human semantics. This provides a verbal option for selecting objects of interest and is potentially, a powerful additional tool for speech interfaces.

3. SYSTEM DESIGN

Our goal is a voice based image parsing system that is simple, fast, and most importantly, intuitive, i.e. allowing an interaction mode similar to our everyday language. After the user loads an image, our system automatically assigns an object class label (nouns) and sets of attributes labels (adjectives) to each pixel. Using these results, our system selects a subset of objects and attributes that are most related to the image. These are shown in Figure 2. These coarse segments provide the bridge between image pixels and verbal commands. Given the various segments, the user can use his/her high level knowledge about the image to strengthen or weaken various object and attribute classes. For example, the initial results in Figure 2 might prompt the user to realize that the bed is missing from the segmentation but the 'cotton' attribute is well defined. Thus, the simple command 'Refine the cotton bed in center-middle' will strengthen the association between cotton and bed, allowing a better segmentation of the bed. Note that the final object boundary does not follow the original attribute segments because verbal information is incorporated as soft cues which are interpreted by a CRF within the context of the other information. Algorithm 1 presents a high level summary of our verbal guided image parsing pipeline, with details explained in the rest of this section.

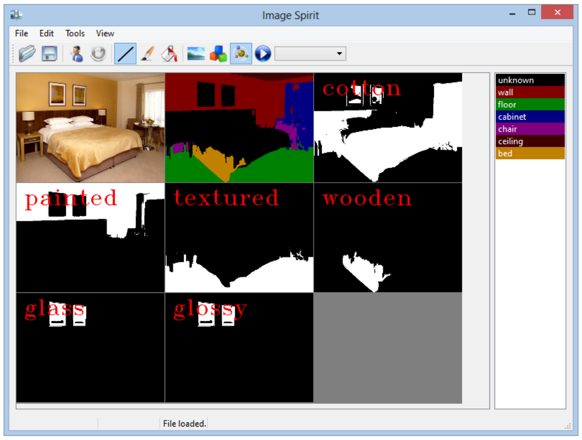


Fig. 2. User interface of our system (labeling thumbnail view).

Once objects have been semantically segmented, it becomes natural to manipulate them using verb-based commands such as move, change, etc. As a demonstration of this concept, we encapsulate a series of rule based image processing commands needed to execute an action, allowing hands-free image manipulation (c.f. Section 5).

3.1 Mathematical Formulation

We formulate simultaneous semantic image segmentation for object class and attributes as a multi-label CRF that encodes both object and attribute classes, and their relations with each other. This is a combinatorially large problem. If each pixel takes one of the 16 object labels and a subset of 8 different attribute labels, there are $(16 \times 2^8)^{640\times480}$ possible solutions to consider for an image of resolution 640 × 480. Direct optimization over such a huge number of variables is currently computational infeasible without some choice of simplification. The problem becomes still more complicated if correlation between attributes and objects are taken into

Algorithm 1 Verbal guided image parsing.

```
Input: an image and learned weak hypothesis (c.f. Figure 1).
Output: an object and a set of attributes labels for each pixel.
Initialize: object/attributes potentials for each pixel as weak hypothesis values; find pairwise potentials by (3).
for Automatic inference iterations $i$ = 1 to 5 do
    Update potentials using (5) and (6) for all pixels simultaneously using efficient filtering technique;
end for
for each verbal input do
    update potentials (c.f. Section 3.3) according to user input;
    for Verbal interaction iterations $i$ = 1 to 3 do
        Update potentials using (5) and (6) as before;
    end for
end for
Get results from potentials: at any stage, labels for each pixel could be found by selecting the largest object potential, or comparing the positive and negative attributes potentials.
```

Table I. List of annotations

Symbols	Explanation (use RV to represent random variable)
$\mathcal{O}$	Set of object labels: $\mathcal{O} = \{o_1, o_2, ..., o_K\}$
$\mathcal{A}$	Set of attribute labels: $\mathcal{A} = \{a_1, a_2, ..., a_M\}$
X_i	A RV for object label of pixel $i \in \{1, 2, ..., N\}$, $X_i \in \mathcal{O}$
$Y_{i,a}$	A RV for attribute $a \in \mathcal{A}$ of pixel i, $Y_{i,a} \in \{0, 1\}$
Y_i	A RV for a set of attributes $\{a : Y_{i,a} = 1\}$ of pixel i
Z_i	A RV $Z_i = (X_i, Y_i)$ of pixel i
$\mathcal{Z}$	RVs of CRF: $\mathcal{Z} = \{Z_1, Z_2, ..., Z_N\}$
z_i, x_i, y_i	Configuration/assignment of RVs Z_i, X_i, Y_i
ψ_i	Unary cost of CRF
ψ_{ij}	Pairwise cost of CRF
$\psi_i^{\mathcal{O}}(x_i)$	Cost of X_i taking value $x_i \in \mathcal{O}$
$\psi_{i,a}^{\mathcal{A}}(y_{i,a})$	Cost of $Y_{i,a}$ taking value $y_{i,a} \in \{0, 1\}$
$\psi_{i,o,a}^{\mathcal{OA}}$	Cost of conflicts between correlated attributes and objects
$\psi_{i,a,a'}^{\mathcal{A}}$	Cost of correlated attributes taking distinct indicators
$\psi_{ij}^{\mathcal{O}}$	Cost of similar pixels with distinct object labels
$\psi_{i,j,a}^{\mathcal{A}}$	Cost of similar pixels with distinct attribute labels

account. In this paper, we propose using a factorial CRF framework [Sutton et al. 2004] to model correlation between objects and attributes.

A multi-label CRF for dense segmentation of objects and attributes can be defined over random variables $\mathcal{Z}$, where each $Z_i = (X_i, Y_i)$ represents object and attributes labels of the corresponding image pixel i (c.f. Table I for a list of annotations). X_i will take value from a set of *object labels* $\mathcal{O}$. Rather than taking values directly in the set of *attribute labels* $\mathcal{A}$, Y_i take values in the *power-set* of the attributes. E.g. for a pixel $y_i = \{wood\}$, $y_i = \{wood, painted, textured\}$, and $y_i = \emptyset$ are all validate assignments. We denote by $\mathbf{z}$ a joint configuration of these random variables, and $\mathbf{I}$ the observed image data. A fully connected multi-label CRF model can be defined as the sum of unary and pairwise cost terms:

$$E(\mathbf{z}) = \sum_i \psi_i(z_i) + \sum_{i<j} \psi_{ij}(z_i, z_j), \tag{1}$$

where i and j are pixel indices that range from 1 to N. The unary cost term $\psi_i(z_i)$ measures the cost of assigning an object label and a set of attributes label to pixel i, considering learned pixel classifiers for both objects and attributes, as well as learned object-attribute and attribute-attribute correlations. The pairwise cost term $\psi_{ij}(z_i, z_j)$ encourages similar and nearby pixels to take consistent labels.

To optimize (1) we break it down into multi-class and binary subproblems using factorial CRF framework [Sutton et al. 2004], while modeling correlations between object and attributes. The unary term can be further split as:

$$\begin{aligned}\psi_i(z_i) \;=\;& \psi_i^{\mathcal{O}}(x_i) + \sum_a \psi_{i,a}^{\mathcal{A}}(y_{i,a}) + \sum_{o,a} \psi_{i,o,a}^{\mathcal{OA}}(x_i, y_{i,a}) \\ &+ \sum_{a \neq a'} \psi_{i,a,a'}^{\mathcal{A}}(y_{i,a}, y_{i,a'})\end{aligned} \tag{2}$$

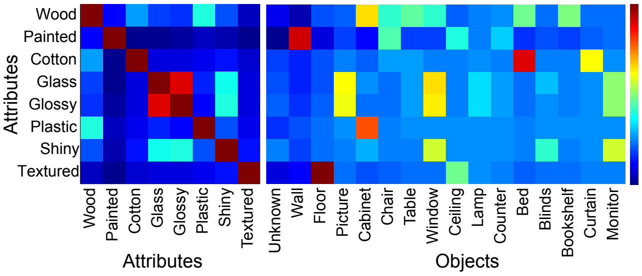


Fig. 3. Visualization of the $R^{\mathcal{OA}}, R^{\mathcal{AA}}$ terms used to encode object-attribute and attribute-attribute relationships.

where the cost of pixel i taking object label x_i is $\psi_i^{\mathcal{O}}(x_i) = -log(\Pr(x_i))$, with probability derived from trained pixel classifier (TextonBoost [Shotton et al. 2009]). For each of the M attributes, we train independent binary TextonBoost classifier, and set $\psi_{i,a}^{\mathcal{A}}(y_{i,a}) = -log(\Pr(y_{i,a}))$ based on the output of this classifier. Finally, the terms $\psi_{i,o,a}^{\mathcal{OA}}(x_i, y_{i,a})$ and $\psi_{i,a,a'}^{\mathcal{A}}(y_{i,a}, y_{i,a'})$ are the costs of correlated objects and attributes with distinct indicators. They are defined as:

$$\psi_{i,o,a}^{\mathcal{OA}}(x_i, y_{i,a}) = [[x_i = o] \neq y_{i,a}] \cdot \lambda_{\mathcal{OA}} R^{\mathcal{OA}}(o, a)$$
$$\psi_{i,a,a'}^{\mathcal{A}}(y_{i,a}, y_{i,a'}) = [y_{i,a} \neq y_{i,a'}] \cdot \lambda_{\mathcal{A}} R^{\mathcal{A}}(a, a')$$

where Iverson bracket, [.], is 1 for a true condition and 0 otherwise, $R^{\mathcal{OA}}(o, a)$ and and $R^{\mathcal{A}}(a, a')$ are derived from learned object-attribute and attribute-attribute correlations respectively. Here $\psi_{i,o,a}^{\mathcal{OA}}(x_i, y_{i,a})$ and $\psi_{i,a,a'}^{\mathcal{A}}(y_{i,a}, y_{i,a'})$ penalize inconsistent object-attributes and attribute-attribute labels by the cost of their correlation value. These correlations are obtained from the ϕ coefficient [Tsoumakas et al. 2009] learnt from the labeled dataset. A visual representation of these correlations are given in Figure 3.

The pairwise cost term $\psi_{ij}(z_i, z_j)$ can be factorized as object label consistent term and attributes label consistent terms:

$$\psi_{ij}(z_i, z_j) = \psi_{ij}^{\mathcal{O}}(x_i, x_j) + \sum_a \psi_{i,j,a}^{\mathcal{A}}(y_{i,a}, y_{j,a}), \tag{3}$$

where the pairwise term takes the form of Potts model [Potts 1952]:

$$\psi_{ij}^{\mathcal{O}}(x_i, x_j) = [x_i \neq x_j] \cdot g(i, j)$$
$$\psi_{i,j,a}^{\mathcal{A}}(y_{i,a}, y_{j,a}) = [y_{i,a} \neq y_{j,a}] \cdot g(i, j)$$

We define $g(i, j)$ in terms of similarity between color vectors I_i, I_j and position values p_i, p_j:

$$\begin{aligned}g(i, j) \;=\;& w_1 \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\mu^2} - \frac{|I_i - I_j|^2}{2\theta_\nu^2}\right) \\ &+ w_2 \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right)\end{aligned} \tag{4}$$

All the parameters $\lambda_{\mathcal{OA}}$, $\lambda_{\mathcal{A}}$, w_1, w_2, θ_μ, θ_ν, and θ_γ are learnt via cross validation. The factorial multi-label CRF model with our verbal guided interaction is illustrated in Figure 4.

3.2 Efficient Joint Inference with Factorized Potentials

An important requirement of our system is that it is near real time, to allow for continuous user interaction. Recently there has been a breakthrough in the mean field solution of random fields, based on recent advances in filtering based methods in computer graphics [Adams et al. 2010; Krähenbühl and Koltun 2011]. Here we briefly sketch how this inference can be extended to multi label CRFs.

This involves finding a mean field approximation $Q(\mathbf{z})$ of the true distribution $P \propto \exp(-E(z))$, by minimizing the KL-divergence $D(Q||P)$ among all distributions Q that can be expressed as a product of independent marginals, $Q(\mathbf{z}) = \prod_i Q_i(z_i)$.

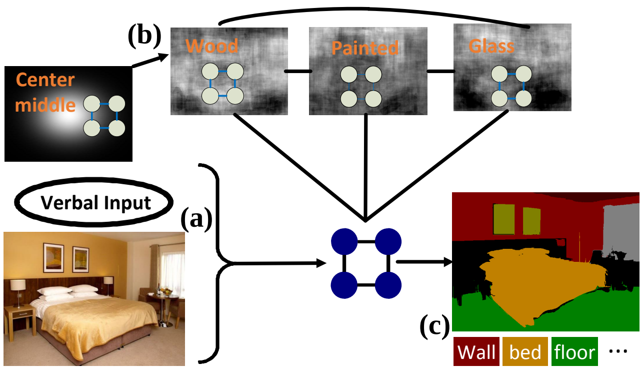


Fig. 4. Verbal guided interactive object class and attributes segmentation in a Factorial CRF framework.

Given the form of our factorial model, we can factorize Q further into a product of marginals over multi-class object and binary attribute variables. Hence we take $Q_i(z_i) = Q_i^{\mathcal{O}}(x_i) \prod_a Q_{i,a}^{\mathcal{A}}(y_{i,a})$, where $Q_i^{\mathcal{O}}$ is a multi-class distribution over the object labels, and $Q_{i,a}^{\mathcal{A}}$ is a binary distribution over $\{0,1\}$.

Given this factorization, we can express the required mean-field updates (c.f. [Koller and Friedman 2009]) as:

$$
\begin{aligned}
Q_i^{\mathcal{O}}(x_i = o) = & \frac{1}{Z_i^{\mathcal{O}}} \exp\{-\psi_i^{\mathcal{O}}(x_i) \\
& - \sum_{i \neq j} Q_j^{\mathcal{O}}(x_j = o)(-g(i,j)) \\
& - \sum_{a \in \mathcal{A}, b \in \{0,1\}} Q_{i,a}^{\mathcal{A}}(y_{i,a} = b)\psi_{i,o,a}^{\mathcal{OA}}(o,b)\} \qquad (5)
\end{aligned}
$$

$$
\begin{aligned}
Q_{i,a}^{\mathcal{A}}(y_{i,a} = b) = & \frac{1}{Z_{i,a}^{\mathcal{A}}} \exp\{-\psi_{i,a}^{\mathcal{A}}(y_{i,a}) \\
& - \sum_{i \neq j} Q_{j,a}^{\mathcal{A}}(y_{j,a} = b)(-g(i,j)) \\
& - \sum_{a' \neq a \in \mathcal{A}, b' \in \{0,1\}} Q_{i,a'}^{\mathcal{A}}(y_{i,a'} = b')\psi_{i,a,a'}^{\mathcal{A}}(b,b') \\
& - \sum_{o} Q_i^{\mathcal{O}}(x_i = o)\psi_{i,o,a}^{\mathcal{OA}}(o,b)\} \qquad (6)
\end{aligned}
$$

where $Z_i^{\mathcal{O}}$ and $Z_{ia}^{\mathcal{A}}$ are per-pixel object and attributes normalization factors. As shown in (5) and (6), directly applying these updates for all pixels requires expensive sum operations, whose computational complexity is quadratic to the number of pixels. Given that our pairwise cost take Potts forms modulated by a linear combination of Gaussian kernels as described in (4), simultaneously finding these sums for all pixels can be achieved at a complexity linear to the number of pixels using efficient filtering techniques [Adams et al. 2010; Krähenbühl and Koltun 2011].

3.3 Attributes/Verbal Interaction

We consider how attributes/verbal interactions can change our unary potentials in (2). Consider the long hypothetical command (in practice we seldom use such long commands) 'Refine the white textured cotton bed in center-middle', The system understands there should be a bed object in the 'center middle'. The 'cotton-textured', 'cotton-bed' and 'textured-bed' correlation matrices should be increased and there should be more weight given to white pixels.

We enforce correlation cues by updating the correlation matrices given in (2). Thus, $\widetilde{R}_{de}^{\mathcal{OA}} = \lambda_1 + \lambda_2 R_{de}^{\mathcal{OA}}$, $\widetilde{R}_{df}^{\mathcal{OA}} = \lambda_1 + \lambda_2 R_{df}^{\mathcal{OA}}$ and $\widetilde{R}_{ef}^{\mathcal{AA}} = \lambda_3 + \lambda_4 R_{ef}^{\mathcal{AA}}$ where d is the bed object index, e the cotton attribute index , f the textured attribute index, λ_i are tuning parameters. and $\widetilde{}$ indicates the updated term.

Location and color information is incorporated by creating a response map $\mathbf{R}$ (c.f. Section 3.4 for how to get color and spatial response maps). We use these response maps to update the corresponding object and attribute unary potentials, $\psi_i^{\mathcal{O}}(x_i), \psi_{i,a}^{\mathcal{A}}(y_{i,a})$ in (2). Specifically, we set

$$\widetilde{\psi}_i^{\mathcal{O}}(x_i) = \psi_i^{\mathcal{O}}(.) - \frac{\lambda_5}{R(i)}, \text{ if } x_i = d \qquad (7)$$

where $\psi_i^{\mathcal{O}}(x_i)$ is unary for objects. Attribute unaries are updated in a similar manner and share the same λ_1 parameter. The $\lambda_{1,..,5}$ parameters are manually set. After these unaries are reset, the inference is re-computed to obtain the updated segmentation result.

3.4 Implementation Details

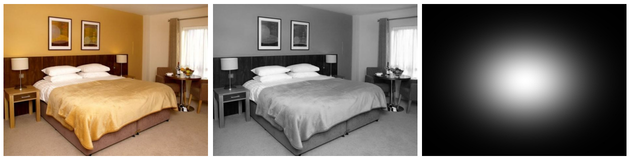
Speech parsing. We use the freely available Microsoft speech SDK [2012] to convert a spoken command into text. Our paper's focus is not on natural language interpretation. Hence, we use a simple speech grammar, with small number of fixed commands. Figure 5 illustrates the 7 commands currently supported.

Supported object classes (**Obj**) include the 16 keywords in our training object class list (unknown, wall, floor, picture, cabinet, chair, table, window, ceiling, lamp, counter, bed, blinds, bookshelf, curtain and monitor). We also support 4 material attributes (**MA**) keywords (wooden, cotton, glass, plastic) and 4 surface attributes (**SA**) keywords (painted, textured, glossy, shiny). For color attributes (**CA**), we support the 11 basic color names, suggested by Linguistic study [Berlin and Kay 1991]. These colors names/attributes include: black, blue, brown, grey, green, orange, pink, purple, red, white and yellow. As also observed by [Laput et al. 2013], humans are not good at describing precise locations but can easily refer to some rough positions in the image. We currently support 9 rough positional attributes (**PA**), by combining 3 vertical

Basic definitions:
MA, **SA**, **CA**, **PA**, are attributes keywords in Section 3.4.
Obj is an object class name keyword in Section 3.4.
ObjDes := [**CA**] [**SA**] [**MA**] **Obj** [in **PA**]
DeformType ∈ {'lower', 'taller', 'smaller', 'larger'}
MoveType ∈ {'down', 'up', 'left', 'right'}

Verbal commands for image parsing:
Refine the **ObjDes**.
Correct the **ObjDes** as **Obj**.
Verbal commands for manipulation:
Activate the **ObjDes**.
Make the **ObjDes DeformType**.
Move the **ObjDes MoveType**.
Repeat the **ObjDes** and move **MoveType**.
Change the **ObjDes** [from **Material/Color**] to **Material/Color**.

Fig. 5. Illustration of supported verbal commands for image parsing and subsequent image manipulation (Section 5). The brackets '[]' represent optional words.



(a) source image (b) white $\mathbf{R}_c$ (c) center-middle $\mathbf{R}_s$

Fig. 6. Response maps of $\mathbf{R}_c$ and $\mathbf{R}_s$ for attributes 'white' and 'center-middle' respectively.

positions (top, center, and bottom) and 3 horizontal positions (left, middle, and right). Since the structure of our verbal commands and the candidate keywords list are fixed, the grammar definition API of Microsoft speech SDK allows us to robustly capture user speech commands. For more sophisticated speech recognition and parsing, we refer the user to use the method in [Laput et al. 2013].

Color $\mathbf{R}_c$ and spatial $\mathbf{R}_s$ attributes response map. Colors are powerful attributes that can significantly improve performance of object classification [van de Sande et al. 2010] and detection [Shahbaz Khan et al. 2012]. To incorporate color into our system, we create a color response map, with the value at ith pixel, $R_c(i)$, defined according to color distance of this pixel to user specified color. We also utilize the location information present in the command to localize objects. Similar to color, the spatial response map value at ith pixel, $R_s(i)$, is defined as the exponent of the negative distance from indicated position. Figure 6 illustrated an example of color position attributes generation according to user speech. The spatial and color response maps are combined into a final overall map $R(i) = R_s(i)R_c(i)$ that is used to update unaries in (7) Since rough color and position names are typically quite inaccurate, we average the initial response values within each region generated by the unsupervised segmentation method [Felzenszwalb and Huttenlocher 2004] for better robustness. These response maps are uniformly scaled to the same range as other object classes' unary potentials for comparable influence to the learned object unary potentials.

Working set selection for efficient interaction. Most images contain only a few of our object classes/ semantic attributes. After the automatic joint object-attribute segmentation stage (takes about 0.5 seconds) we have an approximate idea of the objects and attributes contained in the input image. This allows us to work with a small subset of object classes and attributes during the interactive stage, further increasing system efficiency (0.2 -0.3 seconds) and reducing user ambiguity. We choose this working subset by selecting only those object classes and attributes that have some related pixels in automatic parsing results, or if the class was mentioned by the user in the verbal commands. If users mention an object that does not belongs to any trained object classes, we set the initial unaries of this object class for each image pixel as the average unary of all other classes. The system then interactively segments this object class using the correlated attributes in the user command.

4. EVALUATIONS

aNYU Dataset (attributes augmented NYU). Since per-pixel joint object and attributes segmentation is a new problem, there are only a few existing benchmarks for evaluating it. As also noted by [Tighe and Lazebnik 2011], although the CORE dataset [Farhadi et al. 2010] contains object and attributes labels, each CORE image only contains a single foreground object class, without background annotations. Our focus being more complex complicated images, the CORE dataset is not very suitable for evaluating our method. In order to train our model and perform quantitative evaluation, we augment the widely used NYU indoor V2 dataset [Silberman et al. 2012], through additional labeling of semantic attributes. Figure 7 illustrates an example of ground truth labeling of this dataset. We use the NYU images with ground truth object class labeling, and split the dataset to 724 raining images and 725 testing images. The list of object classes and attributes we use can be found in Section 3.4. We only use the RGB images from the NYU dataset although it provides depth images. Notice that each pixels in the ground truth images are marked with an object class label and a set of attributes labels (on average, 64.7% of them are non empty sets).

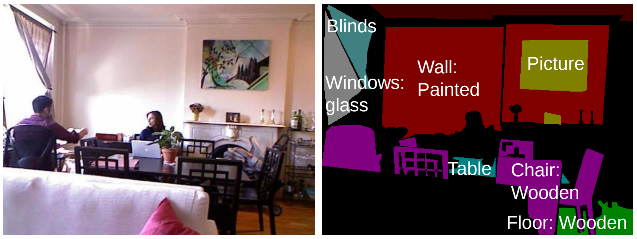


Fig. 7. Example of ground truth labeling in aNYU dataset: original image (left) and object class and attributes labeling (right).

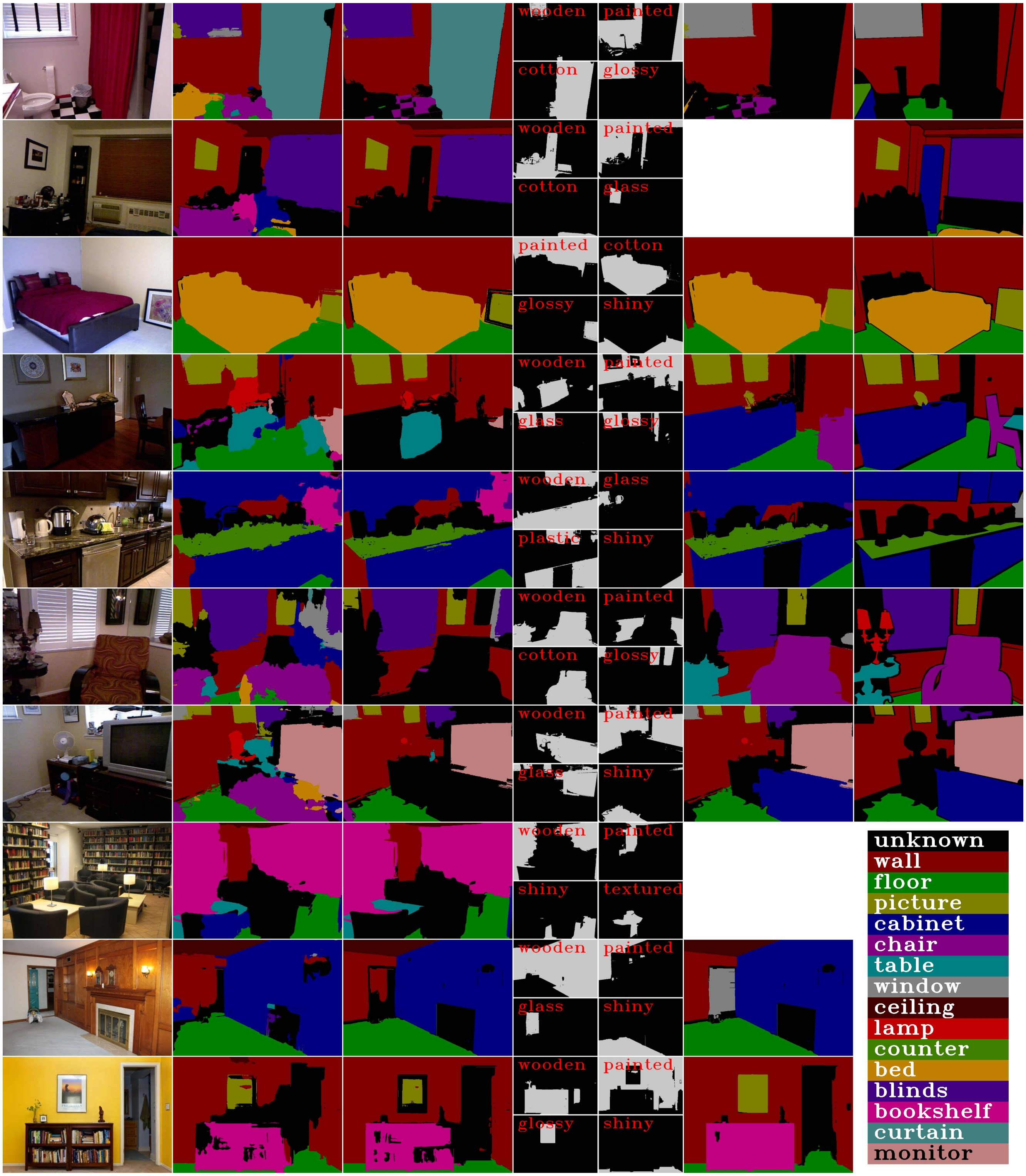
Quantitative evaluation for automatic segmentation. We conduct quantitative evaluation on aNYU dataset. Our approach consists of automatic joint objects-attributes image parsing and verbal guided image parsing. We compared our approach to the state-of-the-art CRF approaches including Associative Hierarchical CRF approach [Ladicky et al. 2009] and Dense CRF [Krähenbühl and Koltun 2011]. Following [Krähenbühl and Koltun 2011], we adopt a label accuracy measure for algorithm performance which is the ratio between number of correctly labeled pixels and total number of pixels. As shown in Table II, our approach significantly outperforms the other state-of-the-art techniques. We have an average accuracy score of 56.6% compared to 50.7% for the previous state of the art.

Quantitative evaluation for verbal guided segmentation. We also performed numerical evaluation for our verbal guided interaction. For this, we choose a subset of 50 images whose collective accuracy scores are reflective of the overall data set. After verbal refinement, our accuracy rises to 80.6% as compared to the $50+\%$ of automated methods. From the results displayed in Figure 8, one

Table II. Quantitative results on aNYU dataset.

Methods	H-CRF	DenseCRF	Our-auto	Our-inter
Label accuracy	51.0%	50.7%	56.9%	- -
Inference time	13.2s	0.13s	0.54s	0.21s
Has attributes	NO	NO	YES	YES

Qualitative results for all 725 testing images can be found in the supplementary. H-CRF (Hierarchical conditional random field model) approach is implemented in a public available library ALE. Dense-CRF represents the state-of-the-art CRF approach. Our-auto stands for our pixel-wise joint objects attributes image parsing approach. Our-inter means our verbal guided image parsing approach. All the experiments are carried out on a computer with Intel Xeon(E) 3.10GHz CPU and 12 GB RAM. Note that all methods in this table use the same features. Without the attributes terms, our CRF formulation will be reduced to exactly the same model as DenseCRF, showing that our JointCRF formulation benefits from the attributes components. Our-inter only considers the time used for updating the previous results given hints from user commands.



(a) source image (b) DenseCRF (c) our object class (d) our attributes (e) verbal refined (f) ground-truth

Fig. 8. Qualitative comparisons. Note that after verbal refinement, our algorithm provides results that correspond closely to human scene understanding. This is also reflected in the numerical results tabulated in Table III. The last three images are from the Internet and lack ground truth. For the second and eight image, there were no attribute combinations which would improve the result, hence there is no verbal refinement.

Table III. Evaluation for verbal guided segmentation.

Methods	DenseCRF	Our-auto	Our-inter
Label accuracy	52.1%	56.2%	80.6%

Here we show average statistics for interacting with a 50 images subset.

can see that these interactive improvements are not just numerical results but also produce object segmentations that accord more to human intuition. In fact, many of the results appear similar to ground truth images! All evaluated images are shown in the supplementary material which bears out this trend.

Note that the final 3 images of Figure 8 (more in supplementary) are not part of the dataset but are Internet images and thus have no ground truth information. These images demonstrate our algorithm's ability to generalize training data for application to images from a similar class (a system trained on indoor images will not work on outdoor scenes) taken under un-controlled circumstances.

User study. Beyond large scale quantitative evaluation, we also test the the plausibility of our new interaction modality by a user study. Our user study comprised of 40 participants, mostly computer science graduates. We study both the time efficiency and the user preference of the verbal interaction. Each user was given a one page instruction script and 1 minute demo video to show how to use verbal commands and mouse tools (line, brush, and fill tool as shown in Figure 2) to interact with the system. After that, each user was given 5 images and asked to improve the parsing results using different interaction modality: i) purely verbal, ii) purely mouse, iii) both of them (in random order to reduce learning bias). Statistics about average interaction time, label accuracy, and user preference is shown in Table IV. In our experiments, participants used a small number of (average 1.7) speech commands to roughly improve the automatic parsing results and then use mouse interaction for further refinements. In this desktop experiment setting, although average preference of verbal interaction is not as good as mouse interaction, it provides a viable alternative to mouse interaction and the combination is preferred by most users. We believe that for new generation of devices such as smart phones and Google glasses, our verbal interaction will be even more useful as it is not easy to perform traditional interactions on them.

Table IV. Comparison of different interaction modality.

Interaction modality	Verbal	Mouse	Verbal + Mouse
Average interaction time (s)	6.9	28.1	11.7
Average label accuracy (%)	80.1	98.1	98.3
Average user preference (%)	12.5	17.5	70.0

Limitations. The limitations of our approach are two fold. Firstly, our reliance on attribute handles can fail if there are no combination of attributes which can be used to improve the image parsing. This can seen in the second and eighth image of Figure 8 where we do not provide any verbal refined result due to lack of appropriate attributes. Of the 78 images we tested (55 from dataset and 23 Internet images) only 10 (5 data-set and 5 Internet images) could not be further refined using attributes. This represents a 13% failure rate. Note that refinement failure does not imply overall failure and the automatic results may still be quite reasonable as seen in Figure 8. Secondly, the vagueness of the language description prevents our algorithm from giving 100% accuracy.

5. MANIPULATION APPLICATIONS

To demonstrate our verbal guided system's applicability as a selection mechanism, we implemented a hands-free image manipulation system. After scene parsing has properly segmented the desired object, we translate the verbs into pre-packaged sets of image manipulation commands. These commands include the in-painting [Barnes et al. 2009] and alpha matting [Levin et al. 2008] needed for a seamless editing effect, as well as semantic specific considerations. The list of commands supported by our system is given in Figure 5 and results in Figure 9. The detailed effects are given below. Note that our hands-free image manipulation is not always fully successful. However, we believe our results are sufficient to demonstrate the possibilities opened up by our verbal scene parsing system.

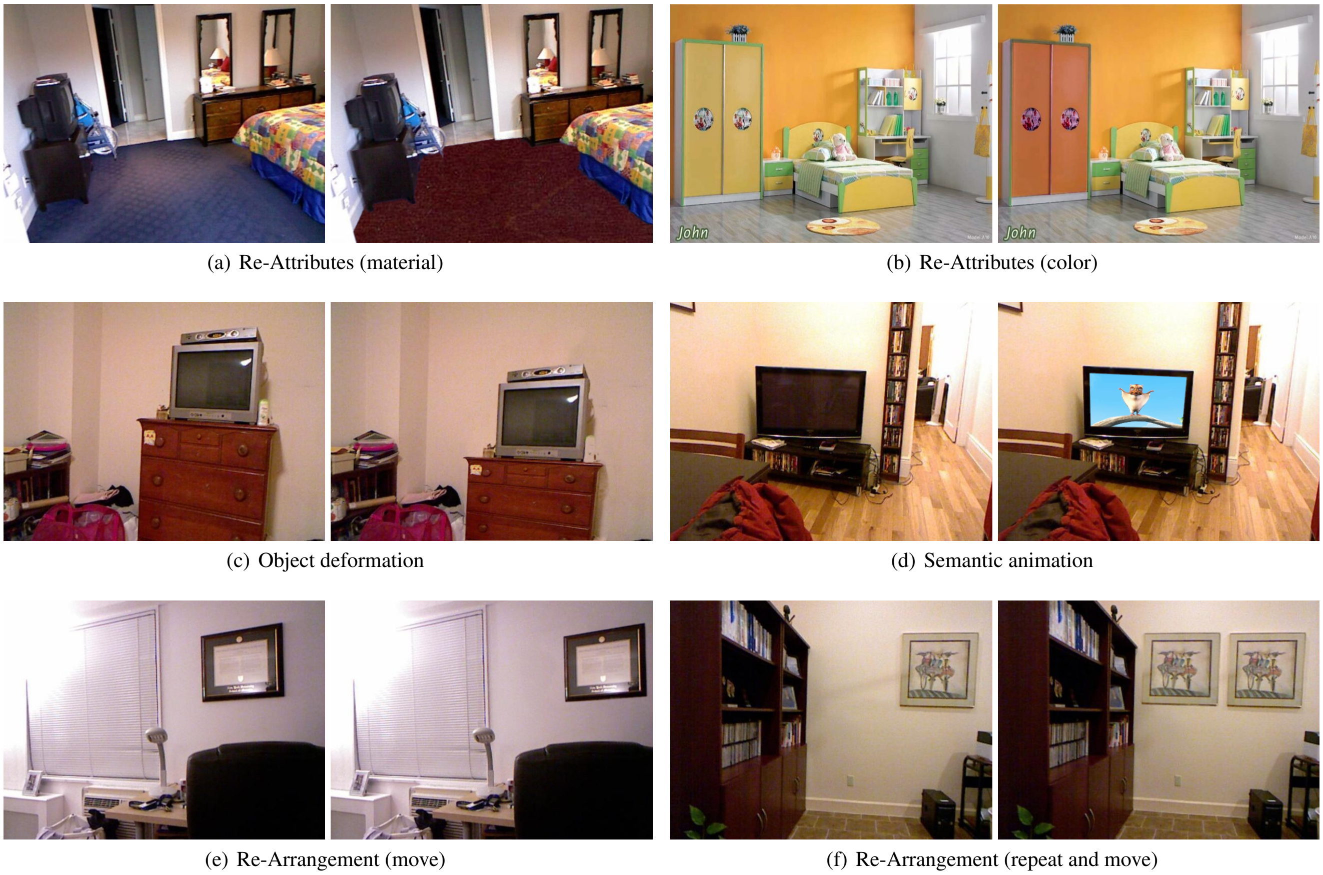
Re-Attributes. Attributes, such as color and surface properties have a large impact on object appearance. Changing these attributes is a typical image edit and naturally lends itself to verbal control. Once the scene has been parsed, one can verbally specify the object to re-attribute. As the computer has a pixel-wise knowledge of the region the user is referring too, it can apply the appropriate image processing operators to alter the region. Some examples are shown in Figure 9. To change object color, we add the difference between average color of this object and the user specified target color. For material changing, we simply tile the target texture (e.g. wooded texture) within the object mask. For more realistic results, we recommend to use texture transfer methods [Efros and Freeman 2001]. Note that in the current implementation, we ignore the surface normal as it is not our target contribution.

Object Deformation and Re-Arrangement. Once an object has been accurately identified, our system supports move, size change and repeat commands which duplicate the object in a new region or changes its shape. In-painting is automatically carried out to refill exposed regions. For greater robustness, we also define a simple, 'gravity' rule for the 'cabinet' and 'table' classes. This requires small objects above these object segments (except stuffs like wall or floor) to follow their motion. Note that without whole image scene parsing, this 'gravity' rule is difficult to implement as there is a concern that a back-ground wall is defined as a small object. Examples of these move commands can be seen in Figure 9, with an example of the 'gravity' constraint in Figure 9c, where the monitor follows the cabinet's motion.

Semantic Animation. Real word objects often have their semantic functions. For example, a monitor could be used to display videos. Since we can get the object region and know its semantic meaning, a natural application would be animating this objects by a set of user or predefined animations. Our system supports an 'activate' command. By way of example consider Figure 9, when user saying 'Activate the black shiny monitor in center-middle', our system automatically fits the monitor region with a rectangle shape, and shows a video in an detected inner rectangle of the full monitor boundary (typically related to screen area). This allows the mimicking real world function of the monitor class.

6. DISCUSSION

This paper presents a novel multi-label CRF formulation for efficient, image parsing into per-pixel object and attribute labels. The attribute labels act as verbal handles through which users can control the CRF, allowing verbal refinement of the image segmentation. Despite the vagueness of verbal descriptors, our system can deliver fairly good image parsing results that correspond to hu-



(a) Re-Attributes (material)

(b) Re-Attributes (color)

(c) Object deformation

(d) Semantic animation

(e) Re-Arrangement (move)

(f) Re-Arrangement (repeat and move)

Fig. 9. Verbal guided image manipulation applications. The commands used are: (a) 'Refine the white wall in bottom-left' and 'Change the floor to wooden', (b) 'Change the yellow wooden cabinet in center-left to brown', (c) 'Refine the glossy monitor' and 'Make the wooden cabinet lower', (d) 'Activate the black shiny monitor in center-middle', (e) 'Move the picture right', (f) 'Repeat the picture in top-left left'. See supplemental video for a live capture of the editing process.

man intuition. Such hands free parsing of an image provides verbal methods to select objects of interest that provides important new iteration modality which enrich the way we interact with images. Both user study and large scale quantitative evaluation verifies the usefulness of our verbal parsing method. Our verbal interaction is especially suitable for new generation devices such as smart phones and Google glasses. To encourage research in this direction, we will release source code and benchmark datasets.

Future work. Possible future directions might include extending our method to video analysis and inclusion of stronger physics based models. Interestingly our system can often segment objects that are not in our initial training set by relying solely on their attribute descriptions. In the future, we might like to carefully select a canonical set of attributes to strengthen this functionality.

REFERENCES

ADAMS, A., BAEK, J., AND DAVIS, M. A. 2010. Fast high-dimensional filtering using the permutohedral lattice. In *Computer Graphics Forum*.

BARNES, C., SHECHTMAN, E., FINKELSTEIN, A., AND GOLDMAN, D. B. 2009. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM TOG* 3, 24:1–11.

BERLIN, B. AND KAY, P. 1991. *Basic color terms: Their universality and evolution*. Univ of California Press.

BLAKE, A., ROTHER, C., BROWN, M., PEREZ, P., AND TORR, P. 2004. Interactive image segmentation using an adaptive gmmrf model. In *ECCV*.

BOLT, R. A. 1980. Put-that-there: Voice and gesture at the graphics interface. In *ACM SIGGRAPH*. 262–270.

BOYKOV, Y. AND JOLLY, M.-P. 2001. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *ICCV*. 105–112.

BRANSON, S., WAH, C., SCHROFF, F., BABENKO, B., WELINDER, P., PERONA, P., AND BELONGIE, S. 2010. Visual recognition with humans in the loop. In *ECCV*. 438–451.

CHEN, T., CHENG, M.-M., TAN, P., SHAMIR, A., AND HU, S.-M. 2009. Sketch2photo: Internet image montage. *ACM TOG 28,* 5, 124:1–10.

EFROS, A. AND FREEMAN, W. 2001. Image quilting for texture synthesis and transfer. *ACM TOG*, 341–346.

FARHADI, A., ENDRES, I., AND HOIEM, D. 2010. Attribute-centric recognition for cross-category generalization. In *CVPR*.

FARHADI, A., ENDRES, I., HOIEM, D., AND FORSYTH, D. 2009. Describing objects by their attributes. In *IEEE CVPR*. 1–8.

FELZENSZWALB, P. AND HUTTENLOCHER, D. 2004. Efficient graph-based image segmentation. *IJCV 59,* 2, 167–181.

FERRARI, V. AND ZISSERMAN, A. 2007. Learning visual attributes. In *NIPS*.

HENDERSON, S. 2008. Augmented Reality for Maintenance and Repair. `http://www.youtube.com/watch?v=mn-zvymlSvk`. [Online; accessed 1-June-2013].

HOSPITAL, S. 2008. Xbox Kinect in the hospital operating room. `http://www.youtube.com/watch?v=f5Ep3oqicVU`. [Online; accessed 1-June-2013].

KOLLER, D. AND FRIEDMAN, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

KRÄHENBÜHL, P. AND KOLTUN, V. 2011. Efficient inference in fully connected CRFs with gaussian edge potentials. In *NIPS*.

KULKARNI, G., PREMRAJ, V., DHAR, S., LI, S., CHOI, Y., BERG, A. C., AND BERG, T. L. 2011. Baby talk: Understanding and generating simple image descriptions. In *IEEE CVPR*. 1601–1608.

LADICKY, L., RUSSELL, C., KOHLI, P., AND TORR, P. H. S. 2009. Associative hierarchical CRFs for object class image segmentation. In *IEEE ICCV*.

LADICKY, L., STURGESS, P., RUSSELL, C., SENGUPTA, S., BASTANLAR, Y., CLOCKSIN, W. F., AND TORR, P. H. S. 2010. Joint optimisation for object class segmentation and dense stereo reconstruction. In *BMVC*.

LALONDE, J., HOIEM, D., EFROS, A., ROTHER, C., WINN, J., AND CRIMINISI, A. 2007. Photo clip art. *ACM TOG 26,* 3, 3.

LAMPERT, C. H., NICKISCH, H., AND HARMELING, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE CVPR*. 951–958.

LAPUT, G., DONTCHEVA, M., WILENSKY, G., CHANG, W., AGARWALA, A., LINDER, J., AND ADAR, E. 2013. Pixeltone: A multimodal interface for image editing.

LEMPITSKY, V., KOHLI, P., ROTHER, C., AND SHARP, T. 2009. Image segmentation with a bounding box prior. In *ICCV*.

LEVIN, A., LISCHINSKI, D., AND WEISS, Y. 2008. A closed-form solution to natural image matting. *IEEE TPAMI*, 228–242.

LI, Y., SUN, J., TANG, C.-K., AND SHUM, H.-Y. 2004. Lazy snapping. *ACM SIGGRAPH 23,* 3, 303–308.

LIU, J., SUN, J., AND SHUM, H.-Y. 2009. Paint selection. *ACM TOG 28,* 3, 1–7.

MICROSOFT. 2012. Microsoft speech platform – SDK. http://www.microsoft.com/download/details.aspx?id=27226.

MORTENSEN, E. AND BARRETT, W. 1995. Intelligent scissors for image composition. In *siggo*. 191–198.

PATTERSON, G. AND HAYS, J. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *IEEE CVPR*. 2751–2758.

POTTS, R. B. 1952. Some generalized order-disorder transformations. In *Proceedings of the Cambridge Philosophical Society*. Vol. 48. 106–109.

RABINOVICH, A., VEDALDI, A., GALLEGUILLOS, C., WIEWIORA, E., AND BELONGIE, S. 2007. Objects in context. In *ICCV*.

ROTHER, C., KOLMOGOROV, V., AND BLAKE, A. 2004. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM TOG 23,* 3, 309–314.

SHAHBAZ KHAN, F., ANWER, R., VAN DE WEIJER, J., BAGDANOV, A., VANRELL, M., AND LOPEZ, A. 2012. Color attributes for object detection. In *IEEE CVPR*. 3306–3313.

SHOTTON, J., WINN, J., ROTHER, C., AND CRIMINISI, A. 2009. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV 81,* 1, 2–23.

SILBERMAN, N., HOIEM, D., KOHLI, P., AND FERGUS, R. 2012. Indoor segmentation and support inference from RGBD images. In *ECCV*.

SUTTON, C., ROHANIMANESH, K., AND MCCALLUM, A. 2004. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In *ICML*.

TIGHE, J. AND LAZEBNIK, S. 2011. Understanding scenes on many levels. In *ICCV*.

TSOUMAKAS, G., DIMOU, A., SPYROMITROS-XIOUFIS, E., MEZARIS, V., KOMPATSIARIS, I., AND VLAHAVAS, I. 2009. Correlation-based pruning of stacked binary relevance models for multi-label learning. In *MLD 2009*.

VAN DE SANDE, K., GEVERS, T., AND SNOEK, C. G. M. 2010. Evaluating color descriptors for object and scene recognition. *IEEE TPAMI 32,* 9, 1582–1596.

VERBEEK, J. AND TRIGGS, W. 2007. Scene segmentation with crfs learned from partially labeled images.

WAH, C., BRANSON, S., PERONA, P., AND BELONGIE, S. 2011. Multiclass recognition and part localization with humans in the loop. In *ICCV*. 2524–2531.

WANG, Y. AND MORI, G. 2010. A discriminative latent model of object classes and attributes. In *ECCV*. 155–168.

ZHENG, Y., CHEN, X., CHENG, M.-M., ZHOU, K., HU, S.-M., AND MITRA, N. J. 2012. Interactive images: Cuboid proxies for smart image manipulation. *ACM TOG 31,* 4, 99:1–11.

ZHOU, S., FU, H., LIU, L., COHEN-OR, D., AND HAN, X. 2010. Parametric reshaping of human bodies in images. *ACM TOG 29,* 4, 126.

Received September 2013; accepted xxx