ImageSpirit: Verbal Guided Image Parsing

Ming-Ming Cheng[1], Shuai Zheng[1*], Wen-Yan Lin[2], Vibhav Vineet[2], Paul Sturgess[2], Nigel Crook[2], Niloy J. Mitra[3], Philip Torr[1]
[1]University of Oxford [2]Oxford Brookes University [3]University College London

Humans describe images in terms of nouns and adjectives while algorithms operate on images represented as sets of pixels. Bridging this gap between how humans would like to access images versus their typical representation is the goal of image parsing, which involves assigning object and attribute labels to pixel. In this paper we propose treating nouns as object labels and adjectives as visual attribute labels. 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. Using the extracted 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, Google Glass, living room devices). We demonstrate our system on a large number of real-world images with varying complexity. To help understand the tradeoffs compared to traditional mouse based interactions, results are reported for both a large scale quantitative evaluation and a user study.

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, natural language control, speech interface

Additional Key Words and Phrases: object class segmentation, image parsing, visual attributes, multi-label CRF

1. INTRODUCTION

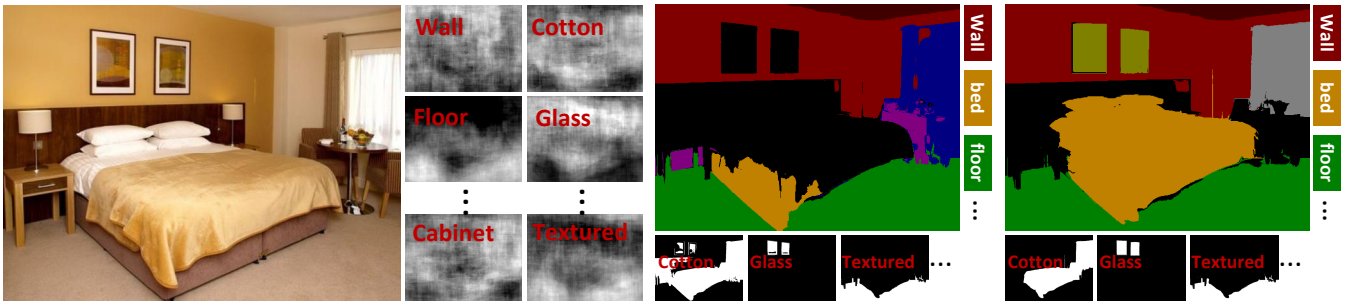
Humans describe images in terms of language components such as nouns (e.g. bed, cupboard, desk) and adjectives (e.g. textured, wooden). In contrast, pixels form a natural representation for computers [Ferrari and Zisserman 2007]. Bridging this gap between our mental models and machine representation is the goal of image parsing [Tu et al. 2005; Tighe and Lazebnik 2013]. The goals of this paper are two-fold: develop a new automatic image parsing model that can handle attributes (adjectives) and objects (nouns), explore how to interact verbally with this parse in order to improve the results. This is a difficult problem. Whilst to date, there exists a large number of automated image parsing techniques [Ladicky et al. 2009; Shotton et al. 2009; Krähenbühl and Koltun 2011; Kulkarni et al. 2011; Tighe and Lazebnik 2011], their parsing results often require additional refinement before being useful for applications such as image editing. In this paper, we propose an efficient approach that allows users to produce high quality image parsing results from verbal commands. Such a scheme enables hands-free parsing of an image into pixel-wise object and attribute labels that are meaningful to both humans and computers. The speech (or speech & touch) input is useful for the new generation of devices such as smart phones, Google Glass, consoles and living room devices, which do not readily accommodate mouse interaction. Such an interaction modality not only enriches how we interact with the images, but also provides an important interaction capability 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 guided[1] image parsing: (i) words are concepts that are difficult to translate into pixel-level meaning; (ii) how to best update the parse using 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 score at each pixel for each object and attribute, e.g. Fig. 1(a). These scores are integrated through a novel, multi-label factorial conditional random field (CRF) model[2] that jointly estimates both object and attribute predictions. We show how to perform inference on this model to obtain an initial scene parse as demonstrated in Fig. 1(b). This joint image parsing with both objects and attributes provides verbal handles to the underlying image which we can now use for further manipulation of the image. Furthermore, our modeling of the symbiotic relation between attributes and objects results in a higher quality parsing than considering each separately [Ladicky et al. 2009; Krähenbühl and Koltun 2011]. To address the second problem, we show how the user commands can be used to update the terms of the CRF. This process of verbal command updating cost, followed by automatic inference to get the results, is repeated until satisfactory results are achieved. Putting the human in the loop allows one to quickly obtain very good results. This is because the user can intuitively leverage a high level understanding of the current image and quickly find discriminative visual attributes to improve scene parsing. For example, in Fig. 1(c), if the verbal command contains the words 'glass picture', our algorithm can re-weight CRF to allow improved parsing of the 'picture' and the 'glass'. Finally, we show

Project page: http://mmcheng.net/imagespirit/. *Joint first author.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2014 ACM 0730-0301/2014/14-ARTXXX $10.00
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.
[2]We substantially extended this model in [Zheng et al. 2014] to include hierarchical relations between regions and pixels, improved attribute-object relationship learning, etc.



(a) Inputs: an image and object/attributes potentials [Shotton et al. 2009] (b) Automatic scene parsing results (c) Natural language guided parsing

Fig. 1. (a) Given a source image downloaded from the Internet, our system generates multiple weak object/attributes cues. (b) Using a novel multi-label CRF, we generate an initial per-pixel object and attribute labeling. (c) The user provides the 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. Best viewed in color. Source image, © Castle Dargan Hotel: http://goo.gl/0cd8V5.

that our joint CRF formulation can be factorized. This permits the use of efficient filtering based techniques [Krähenbühl and Koltun 2011] to perform inference at interactive speed.

We evaluate our approach on the attribute-augmented NYU V2 RGB image dataset [Silberman et al. 2012] that contains 1449 indoor images. We compare our results with state-of-the-art object-based image parsing algorithms [Ladicky et al. 2009; Krähenbühl and Koltun 2011]. We report a 6% improvement in terms of average label accuracy (ALA) [3] using our automated object/attribute image parsing. Beyond these numbers, our algorithm provides critical verbal handles for refinement and subsequent edits leading to a significant improvement (30% ALA) when verbal interaction is allowed. Empirically, we find that our interactive joint image parsing results are better aligned with human perception than those of previous non-interactive approaches, as validated by extensive evaluation results provided in the supplementary material. Further, we find our method performs well on similar scene types taken from outside of our training database. For example, our indoor scene parsing system works on internet images downloaded using 'bedroom' as a search word in Google.

Whilst scene parsing is important in its own right, we believe that our system enables novel human-computer interactions. Specifically, by providing a hands-free selection mechanism 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 verbal guided image manipulation which can be integrated in smart phones and Google Glass, 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 verbal commands 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.

[3]Label accuracy is defined as the number of pixels with correct label divided by the total number of pixels.

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 addressing 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 optimization of stereo and object label [Ladicky et al. 2010], dealing with partial labeling [Verbeek and Triggs 2007], and efficient inference [Krähenbühl and Koltun 2011]. These methods deal only with object labels (noun) and not attributes (adjectives). Visual attributes [Ferrari and Zisserman 2007] and data association [Malisiewicz and Efros 2008], 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]. These works have 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 remains unaware of object boundaries and the obtained object labeling usually spreads over the entire image. We would like to tackle the problem of image parsing with both objects and attributes. This is a very difficult problem as, in contrast to traditional image parsing 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 boxes, which are used to assist the computer in cutting out the desired object from image [Liu et al. 2009; Li 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. Fi-

nal results are achieved via Graph Cut [Boykov and Jolly 2001]. While widely used, these works do not extend naturally to verbal parsing as the more direct scribbles cannot be replaced with vague verbal descriptions such as 'glass'. The second trend in interactive image labeling incorporates a human-in-the-loop [Branson et al. 2010; Wah et al. 2011], which focuses on recognition of image objects rather than image parsing. They resolve ambiguities by interactively asking users to click on the object parts and answer yes/no questions. Our work can be considered a verbal guided human-in-the-loop image parsing. 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 joint CRF model to resolve the ambiguities.

Semantic-based region selection. Manipulation in the semantic space [Berthouzoz et al. 2011] is a powerful tool and there are a number of approaches. An example is Photo Clip Art [Lalonde et al. 2007] which allows users to directly insert new semantic objects into existing images, by retrieving suitable objects from a database. This work has been further extended to sketch based image composition by automatically extracting and selecting suitable salient object candidates [Cheng et al. 2014] from Internet images [Chen et al. 2009; Goldberg et al. 2012]. Carroll *et al.*[2010] enables perspective aware image warps by using user annotated lines as projective constrains. Cheng *et al.* [2010] analyze semantic object regions as well as layer relations according to user input scribble marking, enabling interesting interactions across repeating elements. Zhou *et al.*[2010] proposed to reshape human image regions by fitting an appropriate 3D human model. Zheng *et al.*[2012] partially recover 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 enabling the use of verbal commands to provide 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]). 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. Independently, we have developed a 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 verbal guided 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 (noun) and sets of attribute labels (adjectives) to each pixel. Based on the initial automatical image parsing results, our system identifies a subset of objects and attributes that are most related to the image. In Fig. 2, to speed up the inference in the verbal refinement stage, our system only consider the subset instead of the whole set of object classes and the attribute labels. The initial automatic image parsing results also provide the bridge between image pixels and verbal commands. Given the parse, the user can use his/her knowledge about the image to strengthen or weaken various object and attribute classes. For example, the initial results in Fig. 2 might prompt the user to realize that the bed is missing from the segmentation but the 'cotton' attribute covers a lot of the same area as is covered by the bed in the image. 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 necessarily follow the original boundary of the attribute because verbal information is incorporated only 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.

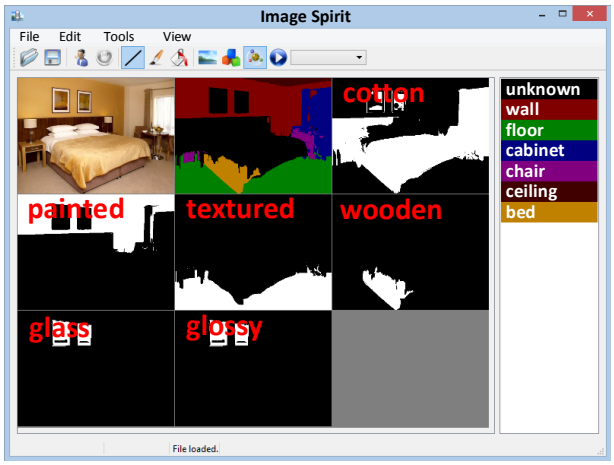


Fig. 2. User interface of our system (labeling thumbnail view).

Once objects have been semantically segmented, it becomes straightforward 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 (see Section 5).

Algorithm 1 Verbal guided image parsing.

Input: an image and object/attributes potentials (see Fig. 1).
Output: an object and a set of attributes labels for each pixel.
Initialize: object/attributes potentials for each pixel; find pairwise potentials by (4).
for Automatic inference iterations i = 1 to T_a **do**
  Update potentials using (6) and (7) 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 T_v **do**
    Update potentials using (6) and (7) as before;
  end for
end for
Extract 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.

3.1 Mathematical Formulation

We formulate simultaneous semantic image parsing for object class and attributes as a multi-label CRF that encodes both object and attribute classes, and their mutual relations. 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 computational infeasible without some choice of simplification. The problem becomes more complicated if correlation between attributes and objects are taken into 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 image parsing of objects and attributes can be defined over random variables $\mathcal{Z}$, where each $Z_i = (X_i, Y_i)$ represents object and attributes variables of the corresponding image pixel i (see Table I for a list of notations). X_i will take a value from the set of object labels, $x_i \in \mathcal{O}$. Rather than taking values directly in the set of attribute labels $\mathcal{A}$, Y_i takes values from the power-set of the attributes. For example, $y_i = \{wood\}$, $y_i = \{wood, painted, textured\}$, and $y_i = \emptyset$ are all valid assignments. We denote by $\mathbf{z}$ a joint configuration of these random variables, and $\mathbf{I}$ the observed image data. Our CRF model is defined as the sum of per pixel and pair of pixel 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 per pixel 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 cost term $\psi_{ij}(z_i, z_j)$ encourages similar and nearby pixels to take similar labels.

To optimize (1) we break it down into multi-class and binary subproblems using a factorial CRF framework [Sutton et al. 2004], whilst maintaining correlations between object and attributes. The pixel term is decomposed into:

Table I. List of notations

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: $\mathcal{A} = \{a_1, a_2, ..., a_M\}$
$\mathcal{P}(\mathcal{A})$	Power set of $\mathcal{A}$: $\mathcal{P}(\mathcal{A}) = \{\{\}, \{a_1\}, ..., \{a_1, ..., a_M\}\}$
X_i	A RV for object label of pixel $i \in \{1, 2, ..., N\}$
$Y_{i,a}$	A RV for attribute $a \in \mathcal{A}$ of pixel i
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: $\mathcal{Z} = \{Z_1, Z_2, ..., Z_N\}$
$y_{i,a}, y_i$	Assignment of RVs $Y_i, Y_{i,a}$: $y_{i,a} \in \{0, 1\}, y_i \in \mathcal{P}(\mathcal{A})$
x_i, z_i	Assignment of RVs X_i, Z_i: $x_i \in \mathcal{O}, z_i = (x_i, y_i)$
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

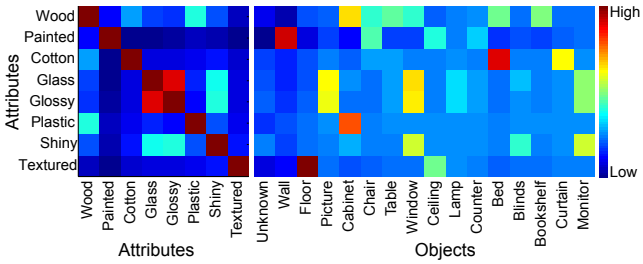


Fig. 3. Visualization of the $R^{\mathcal{OA}}, R^{\mathcal{AA}}$ terms used to encode object-attribute and attribute-attribute relationships.

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

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 classifiers, 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:

$$\begin{aligned} \psi_{i,o,a}^{\mathcal{OA}}(x_i, y_{i,a}) &= [[x_i = o] \neq y_{i,a}] \cdot \lambda_{\mathcal{OA}} R^{\mathcal{OA}}(o, a) \\ \psi_{i,a,a'}^{\mathcal{A}}(y_{i,a}, y_{i,a'}) &= [y_{i,a} \neq y_{i,a'}] \cdot \lambda_{\mathcal{A}} R^{\mathcal{A}}(a, a') \end{aligned} \tag{3}$$

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, which is learnt from the labeled dataset using [Tsoumakas et al. 2009]. A visual representation of these correlations is given in Fig. 3.

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

$$\psi_{ij}(z_i, z_j) = \psi_{ij}^{\mathcal{O}}(x_i, x_j) + \sum_a \psi_{i,j,a}^{\mathcal{A}}(y_{i,a}, y_{j,a}), \tag{4}$$

here we assume each has the form of Potts model [Potts 1952]:

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

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

$$g(i,j) = w_1 \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\mu^2} - \frac{|I_i - I_j|^2}{2\theta_\nu^2}\right) + w_2 \exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right). \tag{5}$$

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.

3.2 Efficient Joint Inference with Factorized Potentials

To enable continuous user interaction, our system must have a response rate which is close to real time. Recently there has been a breakthrough in the mean-field solution of random fields, based on 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, $Q(\mathbf{z}) = \prod_i Q_i(z_i)$. 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:

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

$$\begin{aligned} Q_{i,a}^{\mathcal{A}}(y_{i,a} = b) = & \frac{1}{Z_{i,a}^{\mathcal{A}}} \exp\{-\psi_{i,a}^{\mathcal{A}}(y_{i,a}) \\ & - \sum_{i \neq j} Q_{j,a}^{\mathcal{A}}(y_{j,a} = b)(-g(i,j)) \\ & - \sum_{a' \neq a \in \mathcal{A}, b' \in \{0,1\}} Q_{i,a'}^{\mathcal{A}}(y_{i,a'} = b') \psi_{i,a,a'}^{\mathcal{A}}(b,b') \\ & - \sum_{o} Q_i^{\mathcal{O}}(x_i = o) \psi_{i,o,a}^{\mathcal{OA}}(o,b)\} \end{aligned} \quad (7)$$

where $Z_i^{\mathcal{O}}$ and $Z_{ia}^{\mathcal{A}}$ are per-pixel object and attributes normalization factors. As shown in (6) and (7), directly applying these updates for all pixels requires expensive sum operations, whose computational complexity is quadratic in the number of pixels. Given that our pair of pixel terms are of Potts form modulated by a linear combination of Gaussian kernels as described in (5), simultaneously finding these sums for all pixels can be achieved at a complexity linear in the number of pixels using efficient filtering techniques [Adams et al. 2010; Krähenbühl and Koltun 2011].

3.3 Refine Image Parsing with Verbal Interaction

Since the image parsing results of the automatic approach described in Section 3.1 are still far away from what is our human being perceive from the image and what is required by most image parsing applications such as photo editing, we introduce a verbal interaction modality so that the user can refine the automatic image parsing results by providing a few verbal commands. Each command will alter one of the potentials given in Section 3.1.

Supported object classes (**Obj**) include the 16 keywords in our training object class list (bed, blinds, bookshelf, cabinet, ceiling, chair, counter, curtain, floor, lamp, monitor, picture, table, wall, window and unknown). 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 are: black, blue, brown, grey, green, orange, pink, purple, red, white and yellow. Also as 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 positions (top, center, and bottom) and 3 horizontal positions (left, middle, and right).

Basic definitions:
- **MA**, **SA**, **CA**, **PA**, are attributes keywords in Section 3.3.
- **Obj** is an object class name keyword in Section 3.3.
- **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. 4. Illustration of supported verbal commands for image parsing and manipulation (Section 5). The brackets '[]' represent optional words.

Fig. 4 illustrates the 7 commands that are currently supported. These command can alter the per pixel terms in (2). Notice that both the image parsing commands (e.g. Table II) and the manipulation commands (e.g. Fig. 8) contain object descriptions (**ObjDes**) for verbal refinement. If needed[4], this enables the image parsing to be updated during a manipulation operation. In Fig. 4 the distinction between commands 'refine' and 'correct' is as follows: the former should be given when the label assignment is good but the segment could be better; while, the later is to be given when the label is incorrect.

Consider that user give verbal command 'Refine the **ObjDes**', where **ObjDes=**[**CA**][**SA**][**MA**]**Obj**[in **PA**]'. The system understands there should be a object named **Obj** in the position **PA**, and the correlation cues such as **MA**-**SA**, **MA**-**Obj** and **SA**-**Obj** should be encouraged. We achieve this by updating the correlation matrices given in (3). Thus, the altered object-attribute correlations are changed as $R'^{\mathcal{OA}} = \lambda_1 + \lambda_2 R^{\mathcal{OA}}$ and the modified attribute-attribute correlations are updated as $R'^{\mathcal{A}} = \lambda_3 + \lambda_4 R^{\mathcal{A}}$ where λ_i are tuning parameters.

Speech parsing. We use the freely available Microsoft speech SDK [2012] to convert a spoken command into text. We use a simple speech grammar, with a small number of fixed commands. 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, see [Laput et al. 2013].

[4]When we have perfect image parsing results for the image to be manipulated, we might verbally switch off the function that conducts this combination operation of image parsing and manipulation.

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

Fig. 5. Response maps of $\mathbf{R}_c$ and $\mathbf{R}_s$ for attributes 'white' and 'center-middle' respectively.

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 [Khan et al. 2012]. To incorporate color into our system, we create a color response map, with the value at the ith pixel defined according to the distance between the color of this pixel I_i and a user specified color $\mathbb{I}$. We use $R_c(i) = 1 - \|I_i - \mathbb{I}\|$, where each of the RGB color channels are in the range [0,1]. We also utilize the location information present in the command to localize objects. Similar to color, the spatial response map value at the ith pixel is defined as $R_s(i) = \exp(-\frac{d^2}{2\delta^2})$, where d is the distance between the pixel location and the user indicated position. In the implementation, we use $\delta^2 = 0.04$ with pixel coordinates in both directions normalized to [0,1]. Fig. 5 illustrates an example of color and position attributes generated according to a given verbal command. 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 per pixel terms in (8). 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 Hutenlocher 2004] for better robustness. These response maps are normalized to the same range as other object classes' per pixel terms for comparable influence to the learned object per pixel terms.

We use these response maps to update the corresponding object and attribute per pixel terms, $\psi_i^{\mathcal{O}}(x_i), \psi_{i,a}^{A}(y_{i,a})$ in (2). Specifically, we set

$$\psi'^{\mathcal{O}}_i(x_i) = \psi_i^{\mathcal{O}}(.) - \lambda_5 R(i), \text{ if } x_i = \mathbb{O} \tag{8}$$

where $\psi_i^{\mathcal{O}}(x_i)$ is the per pixel term for objects and $\mathbb{O}$ is the user specified object. Attribute terms are updated in a similar manner and share the same λ_5 parameter. The $\lambda_{1,..,5}$ parameters are set via cross validation. After these per pixel terms are reset, the inference is re-computed to obtain the updated image parsing result.

Working set selection for efficient interaction. Our CRF is factorized for efficient inference over the full set of object and attribute labels. However, since the time it takes to perform inference is dependent on the number of labels that are considered, the interaction may take much longer if there are many labels. To overcome this problem, a smaller working set of labels can be employed during interaction, guaranteeing a smooth user experience. Moreover, as observed in [Sturgess et al. 2012], the actual number of object classes present in an image, is usually much smaller than the total number of object-classes considered (around a maximum of 8 out of 397 in the SUN database [Xiao et al. 2010]). We exploit this observation by deriving the working set as the set of labels in the result of our automatic parsing parse and then updating it as required during interaction, for instance if the user mentions a label currently not in the subset. In our implementation this strategy gives an average timing of around 0.2-0.3 seconds per interaction, independent of the total number of labels considered.

4. EVALUATION

aNYU Dataset (attributes augmented NYU). We created a dataset for our evaluation since per-pixel joint object and attributes segmentation is an emerging problem thus there are only a few existing benchmarks[5]. 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 manual labeling of semantic attributes. Fig. 6 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 into 724 training images and 725 testing images. The list of object classes and attributes we use can be found in Section 3.3. 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).

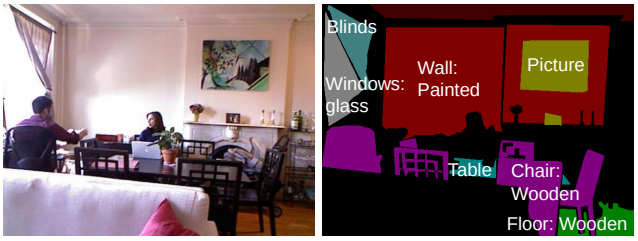


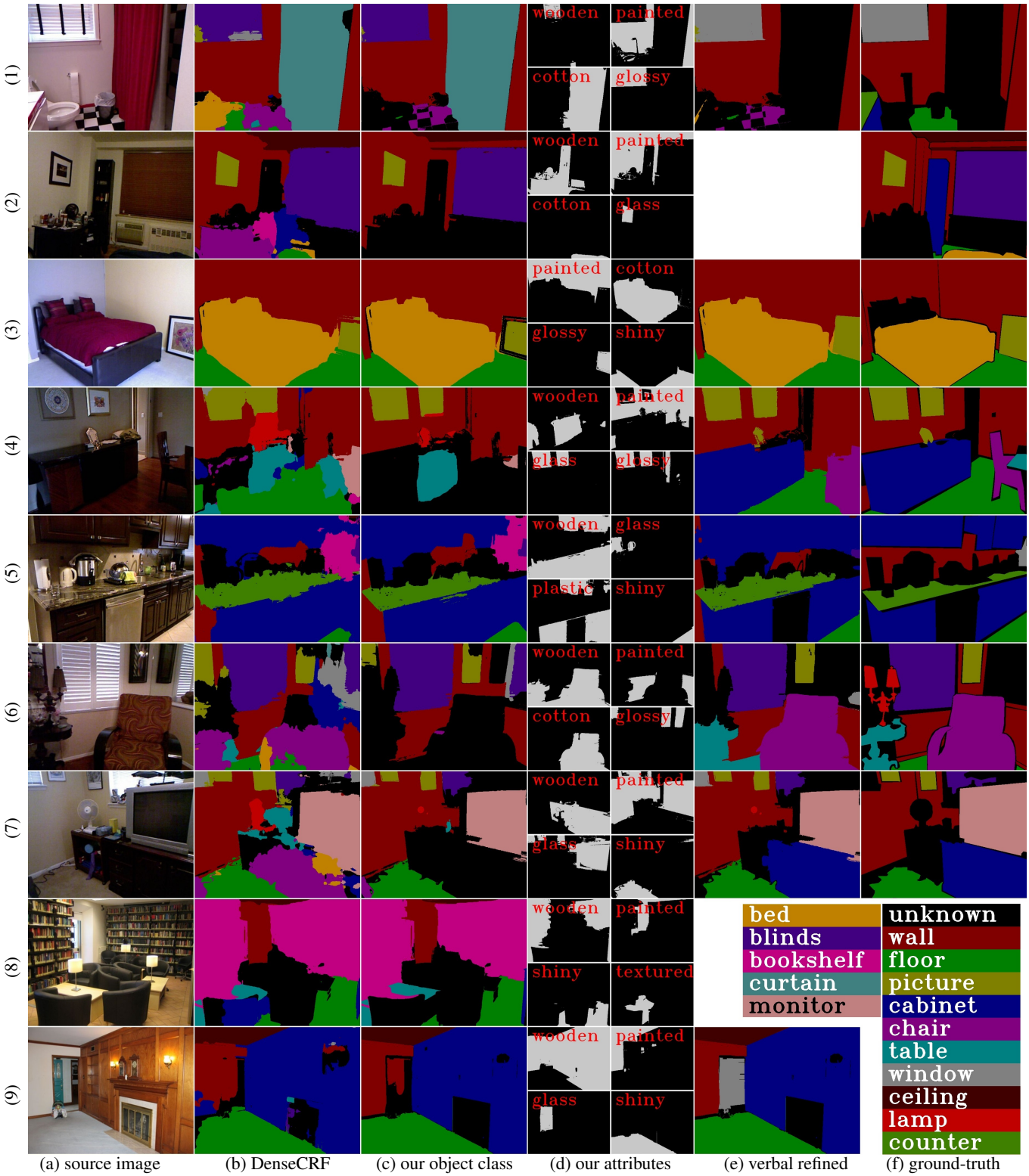
Fig. 6. Example of ground truth labeling in aNYU dataset: original image (left) and object class and attributes labeling (right).

Table II. Verbal commands used for parsing images in Fig. 7.

Image	Verbal commands
(1)	Correct the blinds to window. Correct the curtain to unknown.
(3)	Refine the glossy picture.
(4)	Refine the wooden cabinet in bottom-left. Refine the chair in bottom-right. Refine the floor in bottom-middle.
(5)	Refine the black plastic cabinet. Refine the white unknown in bottom-middle. Refine the cabinet in bottom-left.
(6)	Refine the cotton chair. Refine the glass unknown. Refine the black wooden table in bottom-left.
(7)	Refine the wooden cabinet in bottom-right.
(9)	Refine the glass window.

Quantitative evaluation for automatic image parsing. 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 against two state-of-the-art CRF-based approaches including Associative Hierarchical CRF approach [Ladicky et al. 2009] and Dense CRF [Krähenbühl and Koltun 2011]. For fair comparison, we train the same TextonBoost classifiers for all the methods (a multi-class TextonBoost classifier for object class prediction and M independent binary TextonBoost classifiers, one for each attributes). Following [Krähenbühl and Koltun 2011], we adopt the average label accuracy (ALA) measure for algorithm performance which is the ratio between number of correctly labeled pixels and total number of pixels. As shown in Table III, we have ALA score of 56.6%

[5] 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, without background annotations.



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

Fig. 7. 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 IV. The last three images are from the Internet and lack ground truth. For the second and eight image, there are no attribute combinations which would improve the result, hence there is no verbal refinement. (See Table II for the used verbal commands.) Source image credit, (8) © Wikimedia: (http://goo.gl/KGRP8T), (9) © Associerge (http://goo.gl/O4C2IH), and all other images in this paper are from NYU dataset [Silberman et al. 2012].

Table III. 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 material. The H-CRF (Hierarchical conditional random field model) approach is implemented in a public available library: ALE(`http://cms.brookes.ac.uk/staff/PhilipTorr/ale.htm`), Dense-CRF [Krähenbühl and Koltun 2011] 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.

Table IV. Evaluation for verbal guided image parsing.

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.

compared to 50.7% for the previous state-of-the-art results. During the experiments, we achieve best results when we set $T_a = 5$, as described in Algorithm 1.

Quantitative evaluation for verbal guided image parsing. We numerically evaluate our verbal guided interaction. 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 $50 - 56\%$ of automated methods. From the results displayed in Fig. 7, one can see that these interactive improvements are not just numerical but also produce object segmentations that accord more to human intuition. In fact, many of the results appear similar to ground truth segments. All evaluated images are shown in the supplementary material which bears out this trend. In experiments, we achieve best speed-accuracy-tradeoff results when we set $T_a = 5$, and $T_v = 3$, as described in Algorithm 1.

Note that the final 2 images of Fig. 7 (more in supplementary) are not part of the aNYU dataset but are Internet images without any ground truth annotations. 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 uncontrolled circumstances.

User study. Beyond large scale quantitative evaluation, we also test the plausibility of our new interaction modality by a user study. Our user study comprises of 38 participants, mostly computer science graduates. We investigate 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 Fig. 2) to interact with the system. The users were given 5 images and asked to improve the parsing results using different interaction modality: i) only verbal, ii) only finger touch, iii) both verbal and touch (in random order to reduce learning bias). Statistics about average interaction time, label accuracy, and user preference is shown in Table V. In our experiments, participants use a small number of (mean and standard deviation: 1.6 ± 0.95) verbal commands to roughly improve the automatic parsing results and then touch interaction for further refinements. In the 'verbal+touch' modality,

Table V. Interactive time and accuracy comparison between different interaction modality: verbal, finger touch and both

Interaction modality	verbal	touch	verbal + touch
Average interaction time (s)	6.6	32.3	11.7
Average accuracy (%)	80.3	95.2	97.8
Average user preference (%)	15.8	10.5	73.7

74.4% users preferred verbal command before touch refinement. In desktop setting, although average preference of verbal interaction is not as good as touch interaction, it provides a viable alternative to touch interaction while the combination was generally preferred by most users. We believe that for new generation devices such as Google Glass and other wearable devices, our verbal interaction will be even more useful as it is not easy to perform traditional interactions on them.

5. MANIPULATION APPLICATIONS

To demonstrate our verbal guided system's applicability as a selection mechanism, we implement 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 in-painting [Sun et al. 2005; Barnes et al. 2009] and alpha matting [Levin et al. 2008] needed for a seamless editing effect, as well as semantic rule-based considerations. The list of commands supported by our system is given in Fig. 4 and some sample results in Fig. 8. The detailed effects are given below. Although the hands-free image manipulation results are not entirely satisfactory, we believe that the initial results demonstrate the possibility offered by verbal scene parsing (see also video).

Re-Attributes. Attributes, such as color and surface properties have a large impact on object appearance. Changing these attributes is a common task 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 pixel-wise knowledge of the region the user is referring too, it can apply the appropriate image processing operators to alter it. Among all the pixels with user specified object class label, we choose the 4-connected region with the biggest weight as the extent of the target object, with weights defined by the response map as shown in Fig. 5. Some examples are shown in Fig. 8. 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. wood texture) within the object mask. Alternately, texture transfer methods [Efros and Freeman 2001] can be used. Note that in the current implementation, we ignore effects due to varying surface orientation.

Object Deformation and Re-Arrangement. Once an object has been accurately identified, our system supports move, size change and repeat commands that duplicate the object in a new region or changes its shape. Inpainting is automatically carried out to refill exposed regions. For robustness, we also define a simple, 'gravity' rule for the 'cabinet' and 'table' classes. This requires small objects above these object segments (except stuff such as wall and 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 background wall is defined as a small object. Examples of these move commands can be seen in Fig. 8c.

Semantic Animation. Real word objects often have their semantic functions. For example, a monitor could be used to display videos.

(a) Re-Attributes (material) (b) Object deformation (c) Semantic animation (d) Re-Arrangement (e) Re-Arrangement

Fig. 8. Verbal guided image manipulation applications. The commands used are: (a) 'Refine the white wall in bottom-left' and 'Change the floor to wooden', (b) 'Refine the glossy monitor' and 'Make the wooden cabinet lower', (c) 'Activate the black shiny monitor in center-middle', (d) 'Move the picture right', (e) 'Repeat the picture in top-left and move left'. See supplemental video for a live capture of the editing.

Since we can estimate the object region and its semantic label, a natural manipulation would be animating these objects by a set of user or predefined animations. Our system supports an 'activate' command. By way of example consider Fig. 8, when the user says '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 parsing. Despite the ambiguity of verbal descriptors, our system can deliver fairly good image parsing results that correspond to human intuition. Such hands-free parsing of an image provides verbal methods to select objects of interest, which can then be used to aid image editing. Both the user study and the large scale quantitative evaluation verify the usefulness of our verbal parsing method. Our verbal interaction is especially suitable for new generation devices such as smart phones, Google Glass, consoles and living room devices. To encourage the research in this direction, we will release source code and benchmark datasets.

Limitations. Our approach has some limitations. Firstly, our reliance on attribute handles can fail if there is no combination of attributes that can be used to improve the image parsing. This can seen in the second and eighth image of Fig. 7 where we fail to 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 dataset 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 Fig. 7. Secondly, the ambiguity of language description prevents our algorithm from giving 100% accuracy.

Future work. Possible future directions might include extending our method to video and 3D data [Valentin et al. 2014] analysis and inclusion of stronger physics based models as well as the use of more sophisticated techniques from machine learning. 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 would like to better understand this effect and suitably select a canonical set of attributes to strengthen this functionality. It might also be interesting to explore efficient multi-class object detection algorithms to help working set selection, possibly supporting thousands of object classes [Dean et al. 2013; Cheng et al. 2014]. We have only scratched the surface of verbal guided image parsing with many future possibilities, e.g., how to better combine touch and verbal commands, or how verbal refinement may change the learned models so that they perform better on further refinements.

ACKNOWLEDGMENT

We would like to thank the anonymous associate editor and reviewers for their valuable feedbacks, and Michael Sapienza for the voice over in the video. This research was supported by EPSRC (EP/I001107/1), ERC-2012-AdG 321162-HELIOS, and ERC Starting Grant SmartGeometry 335373.

REFERENCES

ADAMS, A., BAEK, J., AND DAVIS, M. A. 2010. Fast high-dimensional filtering using the permutohedral lattice. *Comput. Graph. 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.

BERTHOUZOZ, F., LI, W., DONTCHEVA, M., AND AGRAWALA, M. 2011. A framework for content-adaptive photo manipulation macros: Application to face, landscape, and global manipulations. *ACM TOG 30,* 5, 120.

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*.

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.

CARROLL, R., AGARWALA, A., AND AGRAWALA, M. 2010. Image warps for artistic perspective manipulation. *ACM TOG 29,* 4, 127.

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.

CHENG, M.-M., MITRA, N. J., HUANG, X., TORR, P., AND HU, S.-M. 2014. Global contrast based salient region detection. *IEEE TPAMI*.

CHENG, M.-M., ZHANG, F.-L., MITRA, N. J., HUANG, X., AND HU, S.-M. 2010. RepFinder: Finding Approximately Repeated Scene Elements for Image Editing. *ACM TOG 29,* 4, 83:1–8.

CHENG, M.-M., ZHANG, Z., LIN, W.-Y., AND TORR, P. H. S. 2014. BING: Binarized normed gradients for objectness estimation at 300fps. In *IEEE CVPR*.

DEAN, T., RUZON, M. A., SEGAL, M., SHLENS, J., VIJAYANARASIMHAN, S., AND YAGNIK, J. 2013. Fast, accurate detection of 100,000 object classes on a single machine. In *IEEE CVPR*.

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. *NIPS*.

GOLDBERG, C., CHEN, T., ZHANG, F.-L., SHAMIR, A., AND HU, S.-M. 2012. Data-driven object manipulation in images. *Comput. Graph. Forum 31*, 265–274.

HENDERSON, S. 2008. Augmented Reality for Maintenance and Repair. `http://www.youtube.com/watch?v=mn-zvymlSvk`.

HOSPITAL, S. 2008. Xbox Kinect in the hospital operating room. `http://www.youtube.com/watch?v=f5Ep3oqicVU`.

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

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 *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 *CVPR*.

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

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.

MALISIEWICZ, T. AND EFROS, A. A. 2008. Recognition by association via learning per-exemplar distances. In *CVPR*.

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

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.

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.

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

STURGESS, P., LADICKY, L., CROOK, N., AND TORR, P. H. S. 2012. Scalable cascade inference for semantic image segmentation. In *BMVC*.

SUN, J., YUAN, L., JIA, J., AND SHUM, H.-Y. 2005. Image completion with structure propagation. *ACM TOG 24,* 3, 861–868.

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*.

TIGHE, J. AND LAZEBNIK, S. 2013. Superparsing-scalable nonparametric image parsing with superpixels. *IJCV 101,* 2, 329–349.

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*.

TU, Z., CHEN, X., YUILLE, A. L., AND ZHU, S.-C. 2005. Image parsing: Unifying segmentation, detection, and recognition. *IJCV 63,* 2, 113–140.

VALENTIN, J., VINEET, V., CHENG, M.-M., KIM, D., IZADI, S., SHOTTON, J., KOHLI, P., NIESSNER, M., CRIMINISI, A., AND TORR, P. 2014. SemanticPaint: Interactive 3d labeling and learning at your fingertips. *ACM TOG*.

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

VERBEEK, J. AND TRIGGS, W. 2007. Scene segmentation with CRFs learned from partially labeled images. In *NIPS*.

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

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

XIAO, J., HAYS, J., EHINGER, K. A., OLIVA, A., AND TORRALBA, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*. 3485–3492.

ZHENG, S., CHENG, M.-M., WARRELL, J., STURGESS, P., VINEET, V., ROTHER, C., AND TORR, P. 2014. Dense semantic image segmentation with objects and attributes. In *IEEE CVPR*.

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.