3-D Volumetric Shape Abstraction from a Single 2-D Image

Pablo Sala
University of Toronto
Toronto, Canada
psala@cs.toronto.edu

Sven Dickinson
University of Toronto
Toronto, Canada
sven@cs.toronto.edu

Abstract

We present a novel approach to recovering the qualitative 3-D part structure from a single 2-D image. We do not assume any knowledge of the objects contained in the scene, but rather assume that they're composed from a user-defined vocabulary of qualitative 3-D volumetric part categories input to the system. Given a set of 2-D part hypotheses recovered from an image, representing projections of the surfaces of the 3-D part categories, our method simultaneously perceptually groups subsets of the 2-D part hypotheses into 3-D part "views", from which the shape and pose parameters of the volumetric parts are recovered. The resulting 3-D parts and their relations offer the potential for a domain-independent, viewpoint-invariant shape indexing mechanism that can help manage the complexity of recognizing an object from a large database.

1. Introduction

In the past 10 years, object recognition has been commonly formulated as object detection, whereby a strong object prior "tests" whether a given configuration of image features satisfies a particular model. However, as the task of object detection gives way to the classical problem of categorizing an unknown object from a large database (with tens of thousands of models), a linear search through the detectors (priors) is intractable. Instead, configurations of causally related features must be formed in a domain-independent manner – the problem of perceptual grouping. When such groups carry enough information, they can be used to query (i.e., index) a large database to yield a small number of promising candidates that might account for the groups. Only then should object priors, corresponding to these candidates, be applied as detectors.

Domain-independent perceptual grouping is necessary but not sufficient for indexing into an object category, for the groups must be abstracted before they can be used as effective indices into a space of generic models. Our previous work [22] recognized the need for a process that abstracts 2-D contour features, bridging the gap between noisy, exemplar-specific contours appearing in an image and salient, categorical contours defining a model. We introduced an image abstraction process that used a collection of abstract 2-D part models (closed contours) to drive both perceptual grouping and shape abstraction, yielding a covering of an image with a set of 2-D abstract part models.

The framework could be seen as a controlled shape "hallucination" process, whereby the coarse shape of a noisy cycle of contours was compared to a given part model. While such abstraction is critical to bridging the gap between actual image features and true categorical features, such hallucination can be highly ambiguous, for the more you're allowed to hallucinate, the more models you can "imagine" from your data. As a result, [22] exhibited good recall but offered poor precision. What was missing was an understanding of the interactions among the shapes, for they are clearly not independent.

In this paper, we extend our previous framework [22] in two very important ways. First, we exploit the interactions of the 2-D shapes to yield a powerful set of constraints that significantly improves precision. By thinking about the problem in 3-D, we instead define the 2-D shapes as the projections of the surfaces of a vocabulary of 3-D qualitative volumetric parts. Moreover, the surface adjacencies of covisible sets of surfaces define aspects[1] over these 2-D shapes [8]. Figure 1(a) illustrates a simple example of a set of three 3-D volumetric part categories, their space of topologically distinct aspects, and the component faces of the aspects. From an input image (Figure 1(b)), we use the 2-D shape vocabulary to generate (using [22]) a low-precision set of face hypotheses, as shown in Figure 1(c). By exploiting the structure of the aspects, we gain a powerful set of constraints with which to select and group the hypotheses, yielding the maximum likelihood solution shown in Figure 1(d).

In our second major contribution, we introduce a novel aspect representation that allows us to learn a mapping from

[1] A 2-D aspect corresponds to a family of topologically equivalent views of an object.

(a) (b) (c) (d) (e)

Figure 1. Overview of the problem: (a) a user-defined vocabulary of 3-D volumetric part categories; by analyzing all possible views over a large set of within-class deformations of the volumetric part categories, we learn a set of projected aspects along with their component faces and relations; (b) original image; (c) using the framework of [22] trained on the component faces of the aspects, a "controlled hallucination" process yields a set of abstract face hypotheses with low precision but high recall; (d) the topological relations (context) between the component faces learned from the aspects provide a set of powerful constraints that are exploited in a probabilistic framework that yields a maximum likelihood "covering" of the image in terms of the faces and aspects derived from the vocabulary, dramatically improving precision; (e) we introduce a new aspect representation that allows us to recover the 3-D shape and pose of a volumetric part from its recovered aspect.

the relative distortions of the 2-D parts (or faces) in a given aspect to the actual 3-D shape and pose of the volumetric part "behind" the aspect, as shown in Figure 1(e). In effect, we extend [22] from a 2-D framework, whereby a vocabulary of 2-D parts is used to drive a 2-D image abstraction process, to a 3-D framework, whereby a vocabulary of 3-D parts (and their 2-D projections) is used to drive a 3-D image abstraction process. The advantage of a 3-D abstraction process is clear. If we can recover a configuration of causally related 3-D parts, including their shapes and poses, then whatever indices we compute over the configuration are viewpoint invariant, supporting a database of object-centered 3-D models, and offering the potential for dramatic space and search complexity savings over view-based object representations.

2. Related Work

The problem of 3-D volumetric part recovery from 2-D and 3-D images has a rich history in computer vision in the 1970's, 1980's, and 1990's, and includes such volumetric abstractions as generalized cylinders [1, 4, 17, 14, 5, 20, 16, 26, 32], superquadrics [24, 18, 25, 10, 6, 12, 7], and geons [8, 3, 21, 2, 19, 27]. For those approaches applied to range data, the problem was well-posed. If one could correctly segment, i.e., partition, the 3-D points into groups representing parts, the fitting of a reduced 3-D model to the points was typically heavily overconstrained, and some success was achieved. However, for the 3-D from 2-D problem, success was typically limited to very simple (sometimes "toy") scenes. The limiting assumption made by this early generation of approaches was that there was a one-to-one correspondence between extracted contours (or regions) in the image and contours (or surfaces) on an abstract part model. To assume that the projected contours and surfaces of an abstract volume explicitly appear in the image constrained scenes to look more like collections of idealized parts rather than real objects. The goal of recovering 3-D volumetric shape from an image was an important one, but the restriction of the early state-of-the-art to simple scenes made it difficult to compete with the emerging appearance-based recognition techniques (in the early 1990's), which could be applied to more ambitious scenes.

There's been an important revival in interest in recovering abstract volumetric parts from a single 2-D image, in support of object recognition, scene understanding, or pose estimation, e.g., [11, 23, 31, 9]. The good news is that unlike their predecessors who worked on simple scenes, these new approaches can deal with much more realistic scenes. Unfortunately, this progress comes at the cost of restricting the space of volumetric deformations (e.g., cuboids/blocks without bending or tapering deformations), restricting pose (e.g., upright cuboids/blocks), or assuming that key features, e.g., corners or occluding boundaries, of the volumes can be locally detected. While this new generation of systems has definitely pushed well past their predecessors, they're still somewhat limited by the same assumption limiting their predecessors: local features, learned or otherwise, are in one-to-one correspondence with abstract model features. In this paper, we present a novel framework that tries to relax some of these assumptions and introduce some new representational ideas that we hope can help the community return to this important problem.

3. Problem Formulation

Our approach begins by hypothesizing (or detecting) a set of abstract 2-D faces representing the projections of 3-D surfaces making up the volumetric parts in the vocabulary. We adopt the abstract face detection approach of [22] which consists of two steps: (1) from a training set of deformed and noise-perturbed instances of faces sampled from a vocabulary of 2-D part models, learn a set of contour classifiers that can be used to efficiently search the oversegmented region boundaries of the image's region adjacency graph for cycles whose coarse shape is similar to one of the model

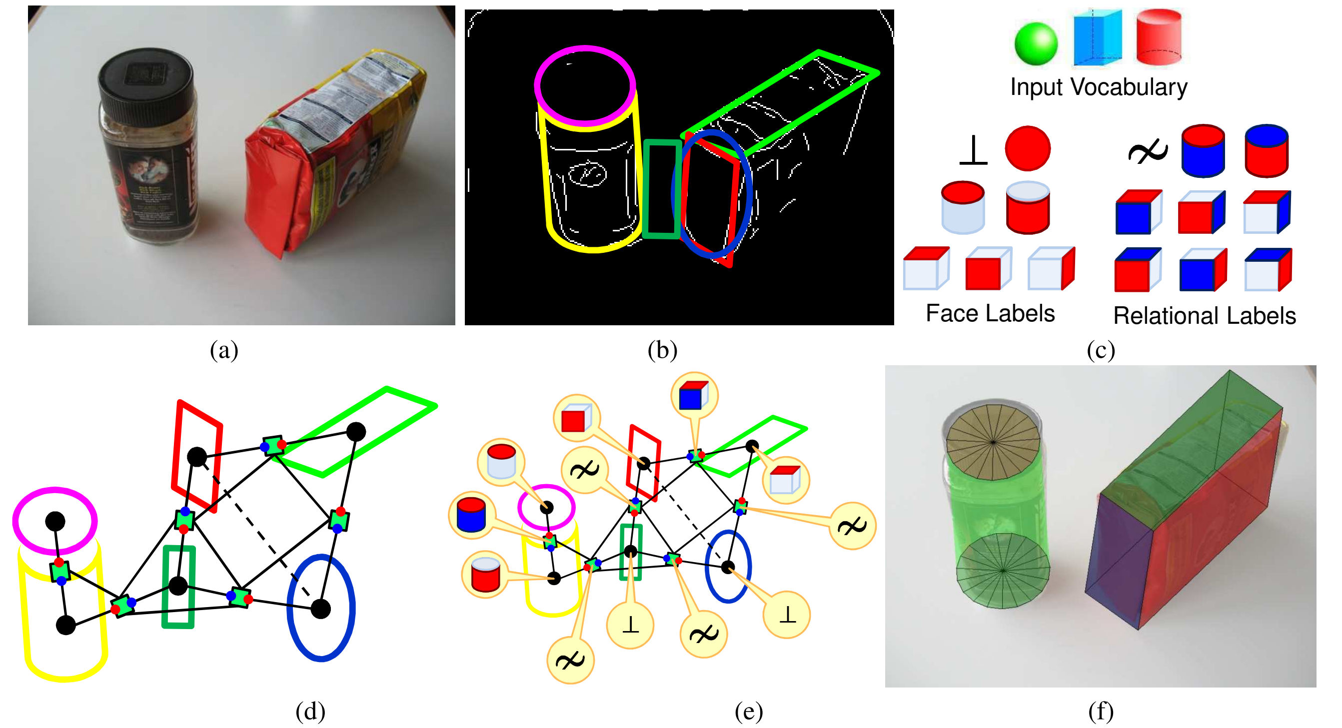


Figure 2. Problem Formulation: (a) input image; (b) output of abstract face detector learned from the projected surfaces of the volumes in the vocabulary shown in (c); note that only a small subset of the detected faces is shown to ease visualization; (d) the face relational graph over the detected faces: black nodes are face nodes, green nodes are relational nodes, solid edges are relational edges, and dashed edges are selection edges; a relational edge between a face node and a relational node represents a pair of proximal faces that might represent adjacent surfaces on a volume, a relational edge between two relational nodes represents a triple of proximal faces that might map to a triple of adjacent surfaces on a volume, and a selection edge spans two nodes that account for the same image evidence, i.e., competing face hypotheses; (e) we seek the maximum likelihood (ML) labeling of the nodes in the graph drawn from the face and relational labels in (c); (f) from the ML labeling, we recover the 3-D shape and pose parameters of the volumes using a novel aspect representation.

parts; and (2) using an active shape model (ASM) trained on the same vocabulary, regularize, i.e., abstract, the shapes of the cycles to yield a set of abstract 2-D part hypotheses. However, unlike [22], which trains their face detectors on a vocabulary of deformed and noise-perturbed 2-D faces, we train our face detectors on the projected surfaces of a vocabulary of deformed and noise-perturbed *3-D volumetric parts* viewed from all possible viewpoints. A simple, illustrative example is shown in Figure 2. From an input image (Figure 2(a)), a set of face hypotheses are detected using the method of [22] (only a small subset of the face hypotheses are shown to ease visualization), as shown in Figure 2(b). The face detectors are trained on the projected surfaces of the deformed and noise-perturbed instances of the volumes shown in Figure 2(c), viewed from all possible viewpoints.

The face hypotheses and their relations are captured in a *face relational graph* containing two types of nodes. A *face node* represents a single face hypothesis (potentially) corresponding to the projection of an abstract volumetric part surface, while a *relational node* represents the grouping of two proximal face hypotheses (potentially) corresponding to the projections of two adjacent surfaces of an abstract volumetric part. Similarly, there are two types of edges in the face relational graph. A *relational edge* connects a pair of nodes (i.e., two relational nodes or a relational node and a face node) that share a common face hypothesis, and is used to enforce the local consistency of the labels of the common face hypothesis across the two nodes that share it. A *selection edge* connects a pair of face nodes whose corresponding face hypotheses have high area and contour overlap, and is used to ensure that competing hypotheses are not simultaneously selected by an interpretation. Returning to our illustrative example, Figure 2(d) depicts the face relational graph derived from the detected face hypotheses in Figure 2(b). Black dots represent face nodes, green squares represent relational nodes, solid lines correspond to relational edges, and dashed lines correspond to selection edges.

Our challenge is to select, from among the low-precision set of face hypotheses in the face relational graph, a set of faces that represents the projected surfaces of volumetric abstractions of the actual 3-D parts that make up the objects in the image. We formulate this as a graph labeling problem in which nodes are labelled according to the set of possible face and relational labels derived from the volumetric part vocabulary (Figure 2(c)). A node's label specifies a 3-D interpretation of the 2-D part hypotheses represented by the (face or relational) node. Let $\mathcal{V}$ be the set of all volumes in the input 3-D part vocabulary, let $\mathcal{A}$ be the set of aspects for all volumes in $\mathcal{V}$, and let $\mathcal{F}$ be the set of faces for all aspects in $\mathcal{A}$. A label is represented as a 3-tuple $(v, a, F) \in \mathcal{V} \times \mathcal{A} \times \mathcal{F}$ indicating a particular volume v from the input 3-D part vocabulary, a specific aspect a of the volume, and a list F of the particular faces in aspect a to be matched to the part hypotheses. In the case of a face node label, $|F| = 1$, and in the case of a relational label, F contains two adjacent faces in the aspect.

Figure 2(c) shows some of the face labels (red) and relational node labels (red-blue) derived from the simple 3-part vocabulary; the labels for the single-face aspects of the cuboid and cylinder and for the two-face aspect of the cuboid are not shown. Note that there is a special face node label $\perp$, which indicates that the node's face hypothesis is not selected by the interpretation, i.e., it is deemed to not correspond to the projected surface of a volume. Similarly, there is a special relational node label $\nsim$, which indicates that the node's face hypotheses are accidentally related, i.e., they do not correspond to the projections of adjacent, co-visible surfaces in a volumetric part. Finally, because a relational label defines two component face labels, we need a mapping from a relational node label to the two face nodes to which it is attached, as shown by the small red and blue dots on the relational (green) nodes in Figure 2(d).

We use a conditional random field (CRF) model to compute the consistent labeling of the graph that yields the 3-D interpretation of the image that best explains the set of 2-D face hypotheses given that they're projections of surfaces of volumetric parts drawn from the vocabulary. We define a probability distribution over clique labels conditioned on the shape and image data of their associated face hypotheses, and we aggregate these conditional clique label probabilities into a global probability model for the entire graph label field. The graph labeling that maximizes this conditional probability over the possible labelings yields a face hypothesis selection (i.e., all surviving face nodes with a label different from $\perp$), a 3-D interpretation for them (indicated by their labels), as well as a surface adjacency interpretation (all surviving relational nodes with a label different from $\nsim$). Returning to our example, Figure 2(e) shows the maximum likelihood labeling of the graph that corresponds to the correct interpretation of the scene.

Finally, in order to efficiently recover the actual pose and parameterization of each volume selected in a labeling, we introduce a novel aspect representation and shape indexing mechanism that learns a mapping between a topological collection of faces (and their shapes) and the surfaces on a particular volume (and its orientation). We are therefore able to infer, from a set of face hypotheses interpreted as a particular aspect, the identity, parameterization, and orientation of the volume whose surfaces project to the faces in the group, as illustrated in Figure 2(f).

4. A Novel Aspect Representation

The aspect of a volumetric part in our vocabulary plays a critical role in our framework. It specifies the relative geometries of the faces making up the aspect, providing a model against which a collection of proximal face hypotheses (extracted form the image) can be compared, i.e., a model which can be used to estimate the conditional probability of a graph clique's label given the face hypotheses. But how do we compare a configuration of face hypotheses extracted from the image, i.e., an *image* aspect, with the configuration of faces that make up a *model* aspect?

We seek a vector representation of an aspect which would not only allow us to model the similarity of two aspects as inversely proportional to the distance between their respective vector representations, but allow us to search a large database of model aspects for the nearest neighbor aspect to the query aspect. Our vector representation must be invariant to translation, planar rotation, and scale. Moreover, the mapping between the faces in an aspect and the vector representation should be distance-preserving in that similar aspects yield similar vectors and vice versa, and continuous in that small perturbations to the relative orientation, translation, and scale between the projected faces in an aspect yield small perturbations in the vector. Finally, we want the mapping to be flexible so that it can be applied to any arbitrary vocabulary of 3-D parts and their aspects.

To meet these representational needs, we introduce a novel aspect representation, called the *aspect signature vector* (ASV), as illustrated for a two-face aspect in Figure 3. Formally, let $F = f_1, \ldots, f_K$ be the list of faces for which an ASV is to be generated ($K \leq 3$). Let d be a direction of face contour traversal (*i.e.*, clockwise or counterclockwise) fixed a priori. Let C_k be an ordered list of T equidistantly sampled points along the 2-D contour of face f_k (in the order resulting from traversing the contour in direction d) and let $C = C_1, \ldots, C_K$. The number T of sampled contour points is fixed a priori and remains the same for all faces and ASVs. ($T = 12$ in the example of Figure 3; in our experiments, we used $T = 64$.) This determines a constant distance δ_k between adjacent sampled contour points within each face f_k. An ASV consists of the rasterization of (a cyclical rotation of) the coordinates of points in $C_1, \ldots, C_K$ (in that order) expressed in a canonical 2-D coordinate system determined by the same set of points. The bottom-right of Figure 3 illustrates the rasterization of one two-face aspect from among its many instantiations shown immediately above.

The origin and orientation of the coordinate system used to compute the ASV is computed from averages of contour

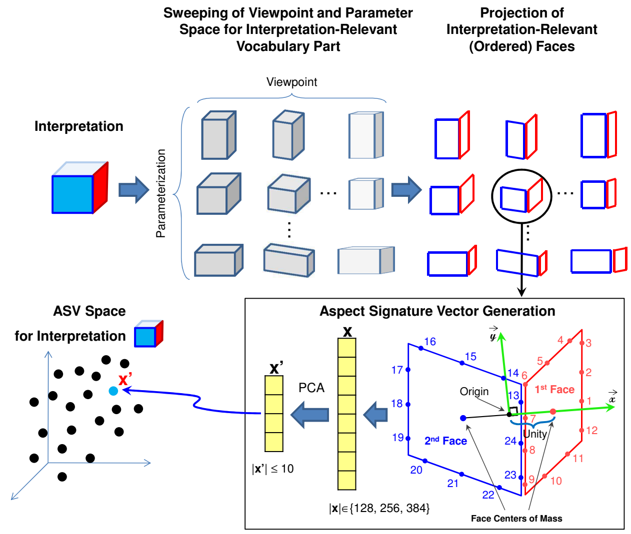


Figure 3. Construction of an ASV space.

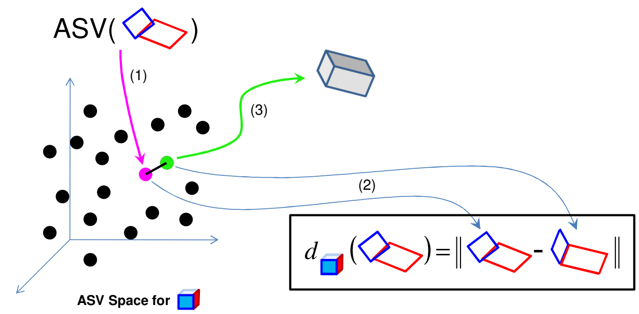


Figure 4. Recovering the 3-D shape and pose of the volume given the aspect label and the ASV of its component faces.

point coordinates, providing robustness to noise. The origin o of the coordinate system is set to the average center of mass of all faces in F, and the coordinate system is oriented such that its x-axis extends from o towards the center of mass of face f_1. The sampled points in each list C_k are cyclically rotated such that the first point in each list in the signature is the one that has the non-negative polar angle (from the face's center of mass in direction $\vec{x}$) closest to zero. (In case of ties, the point closest to the face's center of mass is selected.) Finally, the dimensionality of the ASVs is reduced using PCA, as more than 99% of the total variance of the ASV spaces generated in our implementation is explained by the top ten or fewer components; in more than half of the spaces, three or fewer components were enough. Figure 3 illustrates how the rasterization of an aspect's boundaries leads to an ASV, whose dimensionality is then reduced.

In order to generate the ASV spaces, we discretely sweep the space of parameters and viewpoints of each volume in the vocabulary. Specifically, for each volume in the vocabulary, for each of its parameterizations, for each of its possible aspects, and for each k-face subset of the faces comprising the aspect, (with $k \in \{1, 2, 3\}$), we compute an ASV for each possible 3-D volume orientation that yields the aspect. The ASVs for each distinct face subset are stored in a geometric database, which can be efficiently queried to yield nearest neighbors. Figure 3 shows how, for all possible viewpoints of a cube, the ASVs computed for the 2-face subsets with the label "(cube, aspect with three visible faces, front and side faces)" are stored in a geometric database. We thus end up with a collection of ASV spaces, each representing a different labeling, i.e., 3-D interpretation, of a graph's clique.

We refer to the set of abstract face hypotheses associated to a clique as the clique's *face configuration*. Given a face configuration Q and a clique's label L involving as many faces as $|Q|$, the distance between the ASV of Q and its nearest-neighbor in the ASV space L is called the *interpretation distance of* Q *under* L and is denoted $d_L(Q)$. The set of distances between corresponding contour points on the unrasterization of the ASV of Q and the unrasterization of its nearest-neighbor in the ASV space L is called the *interpretation distance set of* Q *under* L and is denoted $D_L(Q)$. The interpretation distance sets of a clique's face configuration under all possible clique labelings are used as features in the determination of the conditional clique's label probability.

By associating each ASV in a label's ASV space to the shape and pose parameters of the volume used to generate it, we are able to retrieve the closest 3-D model projecting to a given face configuration when that label is assumed as the correct interpretation of the face configuration, as illustrated in Figure 4.

5. Grouping Faces into Aspects

We formulate the problem of grouping face hypotheses into volumetric part aspects as a graph labeling problem in which labels are assigned to the face and relational nodes in the face relational graph. Finding the best volume-aspect interpretation of the image is equivalent to finding the most probable labeling of the graph conditioned on the image evidence (in the form of a collection H of generated 2-D face hypotheses and their associated image data) and the set S of 3-D shape priors induced by the 3-D input vocabulary. We model the graph labeling problem using a conditional random field (CRF), and seek to maximize the conditional probability $p(L|S, H; \Lambda)$ over the label field L.

To keep inference and parameter learning tractable, we consider only cliques of size at most two. There are five possible types of such cliques in the face relational graph, *i.e.*, cliques formed by a single face node, by two face nodes, by a single relational node, by a face node adjacent to a relational node, and by two adjacent relational nodes.

We note these types of cliques as f, ff, r, fr, and rr, respectively. We set clique parameters Λ to depend only on the type of clique and feature function, and so they are the same for all cliques of the same type.

Our probability model thus becomes $p(L|S,H;\Lambda) \propto \exp\left[E(L)\right]$, where

$$E(L) = \sum_{h\in C^f}\left[\overline{\lambda^f}\,\overline{\phi^f}(L_h) \;+\; \lambda^f\,\phi^f(h,L_h)\;\right] + \sum_{h\in C^r}\lambda^r\,\phi^r(h,L_h) + \sum_{h\in C^{rr}}\left[\lambda^{rr}\,\phi^{rr}(h,L_h) \;+\; \psi^{rr}(h,L_h)\right] + \sum_{h\in C^{ff}}\xi^{ff}(L_h) + \sum_{h\in C^{fr}}\psi^{fr}(h,L_h), \quad (1)$$

C^x is the set of face hypothesis configurations of cliques of type x, L_h is the labeling of face hypothesis configuration h as defined by L, and the model parameters are $\Lambda = \left\{\lambda^f, \overline{\lambda}^f, \lambda^r, \lambda^{rr}\right\}$.

We employ four types of feature functions in our model, namely ϕ, $\overline{\phi}$, ξ, and ψ. The most important feature functions are $\phi(L,S,H)$, defined over cliques of type f, r, and rr, that score the quality of each clique's label based on how well the image evidence (configuration of 1, 2, or 3 face hypotheses) matches the model aspect specified by the label (recall that a label is a 3-tuple $(v,a,F) \in \mathcal{V}\times\mathcal{A}\times\mathcal{F}$)). Consider, for example, the top three nodes in the graph in Figure 2(e), namely the face nodes corresponding to the red and lime green faces and the relational node spanning them. Now consider the relational node (clique of type r) and its (correctly shown) label which represents two adjacent surfaces on the cuboid volume. The probability of this label is inversely proportional to the distance between the aspect signature vector of the abstract part hypotheses corresponding to the two face nodes adjacent to the relational node and the aspect signature vector of its closest matching model aspect (from the ASV space of front-and-top-faces in the three-visible-faces aspect of the cuboid model). Figure 4 illustrates how this distance is computed (for a very similar example).

Functions ϕ were implemented as conditional label probabilities via AdaBoost classification, trained on image- and shape-based features from a set of 159 manually annotated images. The features used by these classifiers include: interpretation distance sets (based on the ASVs of the 2-D part hypotheses), which capture the agreement in shape between the part hypotheses and label 3-D interpretations, the generalized boundary measure of Leordeanu *et al*. [13] as a measure of contour saliency, and features assessing the boundary complexity based on variations in curvature along the contours. In the case of relational cliques, the employed features asses the goodness of the contour alignment between the parts, including the distance between closest contour points and the angle between normals at corresponding points.

Features $\overline{\phi}(L)$, defined as binary indicators of type f clique's labels being different from $\perp$, are used to penalize the selection of a large number of hypotheses in a graph interpretation. The roles of the other feature functions are to maintain label consistency. Features $\xi(L_1,L_2)$, defined over type ff cliques (*i.e*., two face nodes connected by a selection edge), prohibit interpretations selecting more than one face hypothesis competing to explain the same image data, by having a value of minus infinity in that case, and zero otherwise. Features $\psi(L,S,H)$ enforce the local label consistency of cliques of types fr and rr by having a value of zero in cases of consistent labelings, and minus infinity otherwise. A labeling of a type fr clique is consistent if either the relational node is labeled $\not\sim$ or if the face node label and relational node label assign the same volume, aspect, and face interpretation to the nodes' common face hypothesis. A labeling of a type rr clique is consistent if either at least one of the two relational node labels is $\not\sim$ or if the face hypothesis common to both nodes is assigned the same volume, aspect, and face interpretation by both relational node labels and the two face hypotheses non-common to both nodes are interpreted as different volume faces.

The generated graphs for the images in our training and test datasets are rather small. The total number of consistent labelings of a graph is highly constrained by its small size, the modest number of relevant interpretations for each node, the large number of constraints yielded by selection edges, and the consistency requirements between adjacent edge labels. Since the resulting number of consistent graph labelings is small, we exhaustively generate all of them via a dynamic programming approach. Perfect inference is thus possible by computing the unnormalized probability (i.e., ignoring the partition function) of each graph labeling and picking the one maximizing it. The small number of graph labelings also makes parameter learning via maximum likelihood estimation simple, since it is possible to compute the partition function in every case.

6. Evaluation

We now demonstrate to what extent our 3-D shape abstraction approach is able to recover the qualitative 3-D shape of an object. Such an evaluation requires a 2-D image dataset in which 3-D objects have been annotated with their compositional abstract volumetric parts, drawn from a vocabulary of qualitative 3-D shape models. To the best of our knowledge, no existing dataset includes ground truth labels, sizes, and poses of the *qualitative* volumes composing the objects in the scene. Datasets with CAD annotations (e.g., [28], [15]) are inappropriate as they do not decompose objects into abstract 3-D parts.[2] Moreover, we are not aware

[2]Note that we are not attempting to recover (or match to) CAD models of entire objects. Rather, our aspects describe a finite vocabulary of ab-

Figure 5. Left: original image; Middle: recovered volumetric part; Right: ground truth.

of any competing 3-D shape abstraction from 2-D approach that has been evaluated on any image dataset.

In the absence of an appropriate dataset, we created an image dataset (which we will make publicly available) containing 100 images of real objects extracted from Caltech 101 and Caltech 256, as well as the Internet.[3] Each image typically contains a non-degenerate view of a single object appearing against a simple background and with no occlusion; a number of objects have some texture and structural detail. The prototypical shape of each selected object can generally be described with a single abstract part. Some of the images in the dataset can be seen in Figures 5 and 7. The images in the dataset were annotated with ground truth abstract volumetric parts (from a given part vocabulary) corresponding to the objects (or object parts) in the scene. The 3-D part vocabulary, used both in our experiments and to annotate the dataset images, consists of cuboids, cylinders, tapered cylinders, ellipsoids and bent ellipsoids.

Qualitative results from our approach are shown in Figure 5. The left column shows the original image. The middle column shows the final output of our system, i.e., the set of 3-D volumes corresponding to the optimal labeling of the face relational graph induced by the set of 2-D part hypotheses generated from each image. Note that the consistent cycles whose abstractions induced the visible faces of the recovered volumes are displayed by dashed lines. The right column shows the annotated ground truth volumetric abstraction(s) for the image.

Notice the ability of our approach to recover volumetric abstractions that do not overlap well with corresponding image contours, e.g., rows 1, 2, 4, 8, 9 in Figure 5. Even in cases where the object's shape is close to that of an ideal volume in the vocabulary, due to image region undersegmentation or noise, the actual detected consistent cycles, which our approach ultimately uses, do not follow well the volume's contours, e.g., rows 3, 7, 10 in Figure 5. In some cases, due to foreshortening or because adjacent object surfaces are not orthogonal, the resulting spatial layout of the face hypotheses does not match well the projected faces of the abstracted volume, e.g., rows 1, 4, in Figure 5. A quantitative evaluation of the proposed approach on the test dataset is displayed in Figure 6, showing the substantial improvement in precision and recall that our method achieves over [22]. (The performance of our approach appears as a single point due to the absence of a natural free parameter in our model that could be varied in order to control the trade-off between precision and recall.)

7. Limitations

There are a number of limitations of our abstraction framework. To begin with, the process of perceptually

stract parts that can be combined (in the spirit of Biederman's RBC theory [3]) to form an infinite number of objects.

[3] Available at http://www.cs.toronto.edu/~psala/datasets.html.

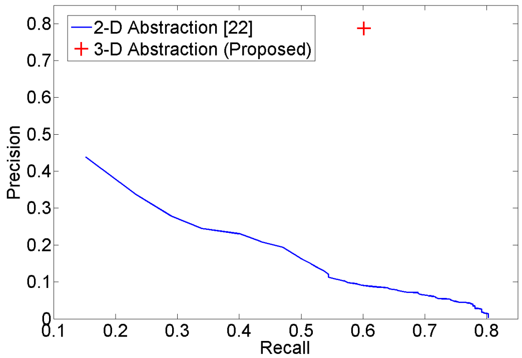
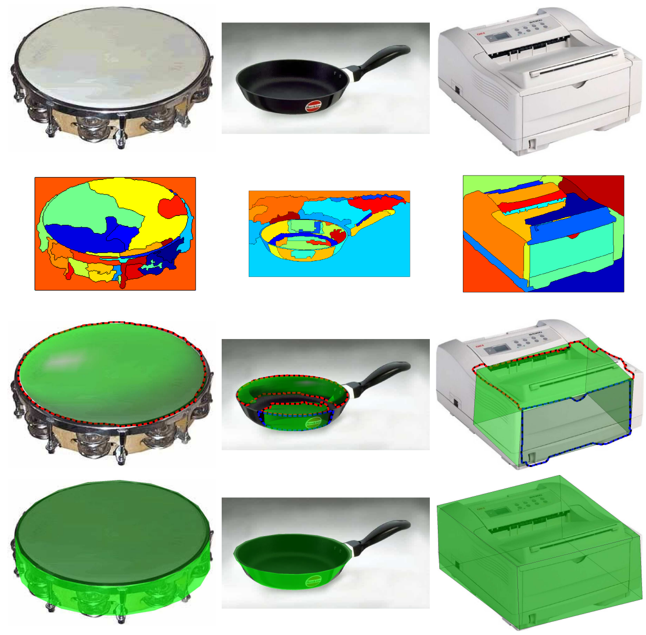


Figure 6. Quantitative Evaluation.

grouping abstract 2-D face hypotheses into aspects (from which volumes are inferred) is limited by the quality of the input set of face hypotheses. For each volumetric part in the scene, we require at least one of its visible 3-D surfaces to be unoccluded and to have a corresponding unoccluded 2-D face hypothesis in the input set.[4] The block in Figure 2 illustrates an example in which only two of the three visible unoccluded surfaces have corresponding face hypotheses (nodes) in the face graph, and the correct volume is still recovered. Our 2-D part hypotheses are generated using the method of [22], which provides a parameter (*abstraction tolerance*) that controls the degree of "hallucination"; setting it to a high value increases face recall at the expense of lowering precision. Their method, in turn, relies on an effective region oversegmentation, with undersegmentation leading to false negatives. Background clutter and rich texture are also problematic for their approach which is based on simple color homogeneity-based region segmentation methods that can yield unwanted oversegmentation which, in turn, significantly increases the complexity of our face relational graph and the resulting labeling problem.

Figure 7 exemplifies some of the limitations of our approach. Input images are shown in the first row, the second row contains the image region segmentations passed as input to [22], the volumetric abstractions generated by our approach are displayed in the third row, and the fourth row shows the ground-truth volumetric abstractions. In (a), a typical error from a false negative 2-D part detection is shown. The recovered volume is incorrect in this case; namely, the elliptical face is interpreted as the projection of an ellipsoid instead of the top face of a cylinder due to the lack of contextual constraints from an undetected adjacent face. In (b), the detection of two false positive volumes results from false positive 2-D parts, while the correct

[4]Unlike the object detection-based approach to aspect layout estimation of Xiang et al. [29, 30], where knowledge of object categories is leveraged, data-driven occlusion reasoning becomes a difficult problem in the absence of categorical conditional feature distributions.



(a) (b) (c)

Figure 7. Limitations (see text for discussion).

2-D parts were not recovered due to region undersegmentation. The shape of the cycle representing the union of regions corresponding to the "side" face of the pan, involving the thin and long curved "arms" surrounding its opening on both sides, was not sufficiently similar to one of the model parts. Finally, in (c), we see an oversegmented volumetric abstraction resulting from an oversegmentation of the top face of the printer, in which a rectangularly-shaped consistent cycle (dashed red) is grouped with the consistent cycle (dashed blue) representing the printer's front face.

8. Conclusions

Our probabilistic formulation of the abstract face hypothesis selection and grouping problem (guided by the face topology encoded in a set of model aspects), our introduction of a new aspect representation that can match a distorted face configuration to a model one, and the resulting shape and pose recovery of the volume "behind" the aspect represent the major contributions of this paper. Our preliminary results indicate that there remains much work on many fronts to overcome the many limitations of the approach. But we strongly believe that as the community returns to the classical problem of recognizing 3-D object categories from a single 2-D image and the classical problem of recovering blocks world-like 3-D parts from a 2-D image, it must rely more on coarse shape and topology, for which we offer a possible path for further exploration.

References

[1] G. Agin and T. O. Binford. Computer description of curved objects. *IEEE Transactions on Computers*, C-25(4):439–449, 1976. 2

[2] R. Bergevin and M. D. Levine. Generic object recognition: Building and matching coarse 3D descriptions from line drawings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:19–36, January 1993. 2

[3] I. Biederman. Human image understanding: Recent research and a theory. *Computer Vision, Graphics, and Image Processing*, 32:29–73, 1985. 2, 7

[4] T. O. Binford. Visual perception by computer. In *Proceedings, IEEE Conference on Systems and Control*, Miami, FL, 1971. 2

[5] R. Brooks. Model-based 3-D interpretations of 2-D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):140–150, 1983. 2

[6] S. Dickinson and D. Metaxas. Integrating qualitative and quantitative shape recovery. *International Journal of Computer Vision*, 13(3):1–20, 1994. 2

[7] S. Dickinson, D. Metaxas, and A. Pentland. The role of model-based segmentation in the recovery of volumetric parts from range data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(3):259–267, March 1997. 2

[8] S. Dickinson, A. Pentland, and A. Rosenfeld. 3-D shape recovery using distributed aspect matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):174–198, 1992. 1, 2

[9] S. Fidler, S. J. Dickinson, and R. Urtasun. 3D object detection and viewpoint estimation with a deformable 3D cuboid model. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Neural Information Processing Systems*, pages 620–628, 2012. 2

[10] A. Gupta and R. Bajcsy. Volumetric segmentation of range images of 3D objects using superquadric models. *CVGIP: Image Understanding*, 58(3):302–326, 1993. 2

[11] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *European Conference on Computer Vision*, 2010. 2

[12] A. Leonardis, A. Jaklic, and F. Solina. Superquadrics for segmenting and modeling range data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(11), 1997. 2

[13] M. Leordeanu, R. Sukthankar, and C. Sminchisescu. Efficient closed-form solution to generalized boundary detection. In A. W. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *European Conference on Computer Vision*, volume 7575 of *Lecture Notes in Computer Science*, pages 516–529. Springer, 2012. 6

[14] D. Marr and H. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Royal Society of London*, B 200:269–294, 1978. 2

[15] K. Matzen and N. Snavely. NYC3DCars: A dataset of 3D vehicles in geographic context. In *IEEE International Conference on Computer Vision*, 2013. 6

[16] G. Medioni and K. Rao. Generalized cones: Useful geometric properties. *Computer Vision, Graphics and Image Processing*, 10(3):185–208, October 1992. 2

[17] R. Nevatia and T. O. Binford. Description and recognition of curved objects. *Artificial Intelligence*, 8:77–98, 1977. 2

[18] A. Pentland. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28:293–331, 1986. 2

[19] M. Pilu and R. B. Fisher. Recognition of geons by parametric deformable contour models. In *Proceedings, European Conference on Computer Vision*, pages 71–82, 1996. 2

[20] J. Ponce. Straight homogeneous generalized cylinders: differential geometry and uniqueness results. *International Journal of Computer Vision*, 4(1):79–100, January 1990. 2

[21] N. Raja and A. Jain. Recognizing geons from superquadrics fitted to range data. *Image and Vision Computing*, 10(3):179–190, April 1992. 2

[22] P. Sala and S. Dickinson. Contour grouping and abstraction using simple part models. In *ECCV*, LNCS 6315, pages 603–616, Crete, Greece, September 2010. 1, 2, 3, 7, 8

[23] J. Schlecht and K. Barnard. Learning models of object structure. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *NIPS*, pages 1615–1623. Curran Associates, Inc., 2009. 2

[24] F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(2):131–146, 1990. 2

[25] D. Terzopoulos and D. Metaxas. Dynamic 3D models with local and global deformations: Deformable superquadrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7):703–714, 1991. 2

[26] F. Ulupinar and R. Nevatia. Shape from contour: Straight homogeneous generalized cylinders and constant cross section generalized cylinders. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(2):120–135, 1995. 2

[27] K. Wu and M. D. Levine. 3-D shape approximation using parametric geons. *Image and Vision Computing*, 15(2):143–158, 1997. 2

[28] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *Workshop on the Applications of Computer Vision* , 2014. 6

[29] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3410–3417, 2012. 8

[30] Y. Xiang and S. Savarese. Object detection by 3D aspectlets and occlusion reasoning. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 530–537, 2013. 8

[31] J. Xiao, B. C. Russell, and A. Torralba. Localizing 3D cuboids in single-view images. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Neural Information Processing Systems*, pages 755–763, 2012. 2

[32] M. Zerroug and R. Nevatia. Volumetric descriptions from a single intensity image. *International Journal of Computer Vision*, 20(1/2):11–42, 1996. 2