A Field Model for Repairing 3D Shapes*

Duc Thanh Nguyen[1,2], Binh-Son Hua[2], Minh-Khoi Tran[2], Quang-Hieu Pham[2], and Sai-Kit Yeung[2]
[1]School of Information Technology, Deakin University, Australia
[2]Singapore University of Technology and Design, Singapore

Abstract

This paper proposes a field model for repairing 3D shapes constructed from multi-view RGB data. Specifically, we represent a 3D shape in a Markov random field (MRF) in which the geometric information is encoded by random binary variables and the appearance information is retrieved from a set of RGB images captured at multiple viewpoints. The local priors in the MRF model capture the local structures of object shapes and are learnt from 3D shape templates using a convolutional deep belief network. Repairing a 3D shape is formulated as the maximum a posteriori (MAP) estimation in the corresponding MRF. Variational mean field approximation technique is adopted for the MAP estimation. The proposed method was evaluated on both artificial data and real data obtained from reconstruction of practical scenes. Experimental results have shown the robustness and efficiency of the proposed method in repairing noisy and incomplete 3D shapes.

1. Introduction

Suppose we are given a set of RGB/RGB-D images of an object captured at multiple viewpoints. The object in the real world (i.e. 3D space) is then re-constructed using some 3D reconstruction algorithm. Ideally, if an object can be observed in RGB/RGB-D images, it can be well reconstructed. However, in reality we have found that the reconstruction often fails even if the RGB/RGB-D data is complete. This is because the matching of the RGB data in structure-from-motion based reconstruction methods (e.g. [14]) could not be done accurately, specially for objects of uniform colours. For reconstruction methods using depth (e.g. [39, 4]), the missing of depth could also cause the incompleteness. We illustrate several cases of this situation in Fig. 1.

Recent advances of 3D acquisition devices and 3D scene reconstruction research [28, 38, 39, 40, 4] have enabled large-scale acquisition of 3D scene data and this has raised a demand on 3D data analysis. However, it often happens that the 3D data cannot be obtained at high quality (as shown in Fig. 1), even by recent reconstruction methods, e.g. [4]. Specifically, the 3D surfaces are missing and/or broken and this phenomenon causes difficulties for many sequential tasks such as 3D object detection and recognition [30, 36], shape analysis [20, 19], and scene understanding [32, 12]. Repairing missing and broken surfaces thus plays a critical role and deserves in-depth study. In this paper, we focus on repairing incomplete 3D shapes. This problem can be also referred to as shape completion. We assume objects are not occluded, i.e. they can be fully observed in RGB/RGB-D images. However, this assumption does not mean that objects can be completely reconstructed.

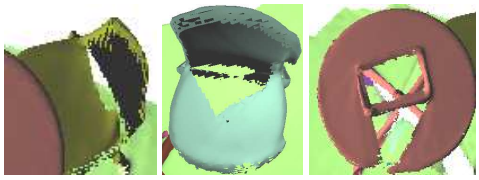

Figure 1. Examples of incomplete shapes after reconstruction using [4].

1.1. Related Work

Existing shape completion approaches make use of geometric information represented at either low-level or high-level. Low-level geometry describes local structures, e.g. local smoothness, and can be used to fill small holes on broken surfaces. For example, Curless and Levoy [5] proposed to extract surfaces by examining the boundary of unseen and empty voxels. However, this method requires additional range images to carve away redundant surfaces. In [7], Davis et al. filled gaps and holes on broken surfaces by performing a convolution on the signed distance values. This process was repeated until a new implicit surface could be defined at the gaps. In [16], a broken object was represented in an octree grid on which inner and outer grid points were determined. The broken object was then constructed by contouring the grid points. In [29], holes on a broken

*This work was conducted when Duc Thanh Nguyen was working at the Singapore University of Technology and Design.

object were filled by local patches (on the same object) best suiting to be pasted at the holes. This method implicitly assumed there were local structures similar to the missing parts at holes. In [17], Kazhdan et al. proposed to interpolate surfaces by fitting surfaces with gradients that could be transformed into a continuous vector field in 3D. This method was then extended in [18] in which constraints at the location of 3D points were incorporated in construction of surfaces. In general, methods using low-level geometric features solely rely on the smoothness constraint, they could potentially resolve small gaps but dare not able to recover large missing parts.

High-level geometric information can be represented via 3D object models. The object models can be predefined using CAD model databases, e.g. [26, 10, 31]. Alternatively, the object models can be constructed based on 2D image segmentation, e.g. [27], or 2D object detection. For example, in [6], image-based object detection method [8] was used to detect the 2D image of an object model, the 3D model was then computed and the pose was estimated. The model was then incorporated into the SLAM system. Recently, Wu et al. [36] proposed to learn the shape model via a convolutional deep belief network. The network was trained on a huge training set of 3D CAD models and then used to recognise and complete broken shapes. Although object-based knowledge could show more advantages compared with low-level geometry information, existing methods of this approach use only 3D models to recover incomplete shapes. This manner holds two limitations. First, the shape repaired using this approach is formed by the predefined/trained 3D models and thus may not represent the real data. Second, RGB data contains rich and useful information (i.e. multi-view data) but is not exploited in shape completion.

1.2. Contributions

To overcome the above issues, we propose a robust and efficient shape repairing method integrating both geometry and multi-view appearance information. In particular, the contributions of the paper include,

- We propose a Markov Random Field (MRF) model for representing 3D shapes. The pairwise priors in the MRF model capture local geometrical structures of the shape and can be learnt using a convolutional deep belief network. The likelihoods are constructed from the multi-view RGB data which is used to verify the consistency of 3D points in various viewpoints.

- We propose a new formulation of shape repairing via maximum a posteriori (MAP) estimation in the MRF model and an efficient inference method for MAP estimation using variational mean field approximation.

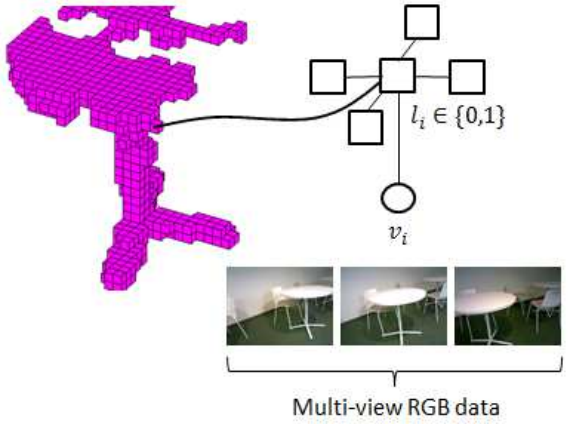

Figure 2. Proposed MRF model.

- We benchmark a new 3D object dataset including objects present in different levels of incompleteness. Compared with existing datasets, e.g. 3D warehouse, SUN database [37], our dataset is more enriched. It includes the 3D models, 2D images captured at various viewpoints, and the 2D-3D correspondences. We will release the dataset to the public as an effort to advance the future research.

The remainder of the paper is organised as follows. Section 2 presents the MRF model and formulates the problem of shape completion. The variational method used for approximation of the MAP estimation is then presented in Section 3. Experimental results and comparisons are reported in section 4. Section 5 concludes the paper with remarks.

2. Problem Formulation

Let S be a 3D shape reconstructed from a sequence of images captured at multiple viewpoints. The shape S can be broken (due to missing data) and/or rough (due to the alignment error of the 3D reconstruction method). We represent the geometric information of the shape S in a 3D voxel grid (see Fig. 2).

Let $G(V, E)$ denote an MRF in which V is the set of voxels in the grid and and E is the set of edges connecting the voxels. Each voxel $v_i \in V$ is associated with a label $l_i \in \{0, 1\}$, $l_i = 1$ if v_i is a voxel of S and $l_i = 0$, otherwise. Similarly to the lattice structure often used in MRFs for 2D image segmentation [2], for each v_i, we consider a set of its 4-connected voxels in the voxel grid; there are 6 neighbours of each voxel (i.e. 4-connected voxels of v_i in a $3 \times 3 \times 3$ cube centred at v_i). The label node l_i of v_i is then connected to the label nodes of the 4-connected voxels of v_i. Fig. 2 illustrate the proposed MRF model.

Let $L = \{l_i\} \in \{0, 1\}^{|V|}$ be a set of labels. The configu-

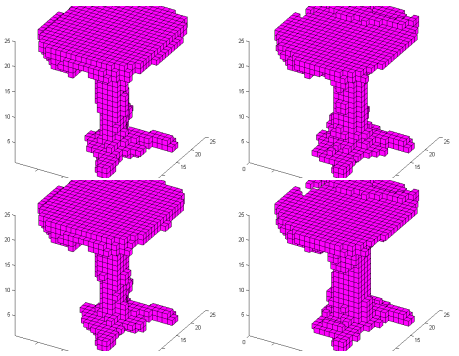

Figure 3. Results of the CDBN applied on the broken table presented in Fig. 2.

ration L is object-specific and modelled via the probability $p(L)$ (also called prior). As conventionally defined in binary MRFs, $p(L)$ can be expressed in the form of Gibbs distribution as

$$p(L) = \frac{1}{Z} \prod_{(v_i,v_j)\in E} \psi_{i,j}(l_i, l_j) \prod_{v_i \in V} \psi_i(l_i) \qquad (1)$$

where $\psi_{i,j}$ and ψ_i are respectively the potentials functions and Z is a normalisation factor (partition function).

In [35], the prior $p(L)$ was used to model a single object class (e.g. pedestrian). However, a single prior is weak to model multiple object classes. In our problem, we construct a multi-model prior $p(L_m)$ representing different object types m (e.g. furniture objects). Motivated by the robustness of deep structures in multi-class object recognition [11], the convolutional deep belief network (CDBN) proposed in [36] is adopted to construct the shape prior $p(L_m)$ in our MRF. The CDBN is enriched by learning from a large set of 3D computer graphics CAD models and thus able to cover many possible object classes. In particular, the reconstructed shape S is fed through the pre-trained CDBN to retrieve a set $\mathcal{M}$ of matching models. Note that the set $\mathcal{M}$ may have more than one model since S is not complete and/or may not exactly match with a unique model. For example, the missing parts of a shape S may be due to the misalignment between image frames used to construct S and those missing parts may be replaced by different parts from different models due to the variation of S. However, each retrieved model $m \in \mathcal{M}$ is a complete shape in which unobserved voxels are predicted by the CDBN trained on various 3D CAD models. Fig. 3 shows several results of the CDBN applied on the broken table presented in Fig. 2.

To make the models m adaptive to small variations of the true shape, we extend the 2D Distance Transform in [9] to the 3D domain and apply it on the models m. Fig. 4 shows two 2D slices across the 3D Distance Transform computed on a result of the CDBN in Fig. 3. By using the Distance Transform, the prior $p(L_m)$ is not restricted to the models m but allows an extent of m. This idea is similar to the 2D shape band proposed in [1]. In particular, we define,

$$\psi_{m,i,j}(l_i, l_j) \propto \exp\left[\alpha f(D_m(i), D_m(j)) l_i l_j\right] \qquad (2)$$

where $\alpha > 0$ is a user-defined parameter, $D_m(i)$ is the value of the Distance Transform of the model m at voxel v_i, and $f(D_m(i), D_m(j))$ is some activation function representing the co-occurrence of l_i and l_j. A low value of $D_m(i)$ indicates that v_i is close to m and vice versa.

The activation function $f(D_m(i), D_m(j))$ is defined as,

$$f(D_m(i), D_m(j)) = \frac{-1}{1 + e^{-\sqrt{D_m^2(i) + D_m^2(j)}}} + \epsilon \qquad (3)$$

where $0.5 < \epsilon < 1$.

The function $f(D_m(i), D_m(j))$ is used to regulate the pairwise prior $\psi_{m,i,j}(l_i, l_j)$ in (2) in the principle that locations close to the model m should have higher value of $\psi_{m,i,j}(l_i, l_j)$ than ones far from m. Indeed, if v_i and v_j are truly empty voxels, $D_m(i)$ and $D_m(j)$ would have high value and $\frac{-1}{1+e^{-\sqrt{D_m^2(i)+D_m^2(j)}}} \to -1$. Thus, $f(D_m(i), D_m(j)) < 0$, i.e. $\psi_{m,i,j}(1,1) < 1 = \psi_{m,i,j}(0,0)$. In other words, $\psi_{m,i,j}(l_i, l_j)$ would attain high value when l_i and l_j are considered as empty voxels. Note that when $l_i = l_j = 1$, $\sqrt{D_m^2(i) + D_m^2(j)}$ in (3) is a variant of the Mahalanobis distance in which the covariance matrix is diagonal and the standard deviation is set to 1.

In contrast, if v_i and v_j are close to m (or even if they are the voxels of m), $D_m(i)$ and $D_m(j)$ would tend to 0 and $f(D_m(i), D_m(j))$ would become > 0, i.e. $\psi_{m,i,j}(1,1) > 1 = \psi_{m,i,j}(0,0)$. In other words, the prior in this case is in favour of considering v_i and v_j as foreground voxels.

In a similar way, we define the potential $\psi_{m,i}$ in (1) through an activation function $g(D_m(i))$ as follows,

$$\psi_{m,i}(l_i) \propto \exp\left[\beta g(D_m(i)) l_i\right] \qquad (4)$$

where $\beta > 0$ and

$$g(D_m(i)) = \frac{-1}{1 + e^{-D_m(i)}} + \epsilon \qquad (5)$$

As in conventional MRFs, the likelihood functions $p(v_i|l_i)$ can be computed based on the observation data which is the RGB images in our case. Specifically, the likelihoods $p(v_i|l_i)$ are defined based on the consistency of the image appearance observed at different viewpoints as follows. Let $\mathcal{I}_i = \{I_1, I_2, ...\}$ be the set of images on which v_i can be observed. Assume that the images in $\mathcal{I}_i$ are ordered in time and the difference in the camera's tilt of two adjacent images I_j and I_{j+1} is about an angle θ. Such sets $\mathcal{I}_i$ can be determined given the temporal sequence of input frames and the camera pose estimated during the reconstruction.

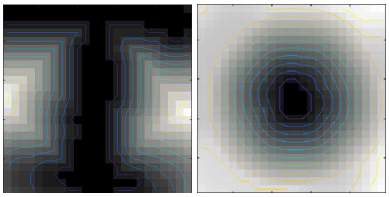

Figure 4. 2D slices across the 3D Distance Transform computed on a result of the CDBN in Fig. 3. Darker values represent small distance and vice versa.

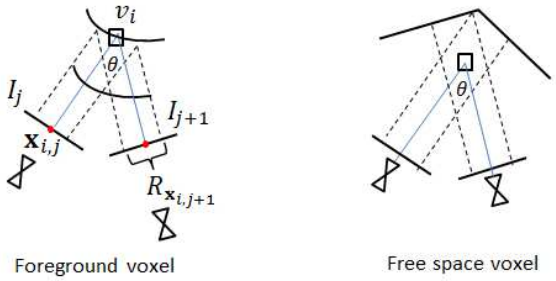

Figure 5. Illustration of likelihood computation for a foreground voxel (left) and free space voxel (right).

Let $\mathbf{x}_{i,j}$ denote the corresponding pixel of v_i on an image $I_j \in \mathcal{I}_i$. A local image region $R(\mathbf{x}_{i,j})$ centred at $\mathbf{x}_{i,j}$ on I_j is then determined. On $R(\mathbf{x}_{i,j})$, we extract a histogram of oriented gradients (HOG) [3], denoted as $h_{R(\mathbf{x}_{i,j})}$. The HOG captures the local appearance of the region $R(\mathbf{x}_{i,j})$. It is expected that if v_i is a foreground voxel, then the HOGs extracted at regions $R(\mathbf{x}_{i,j})$ and $R(\mathbf{x}_{i,j+1})$ on adjacent frames I_j and I_{j+1} should be consistent. However, to achieve this, adjacent frames I_j and I_{j+1} need to be sampled so that they are not very far yet not too close to each other. This is because, when I_j and I_{j+1} are too far from each other, a foreground voxel would even have quite different HOGs on those frames. On the other hand, if I_j and I_{j+1} are too close, an empty voxel even would have very similar HOGs on those frames. In our experiment, we use an angle θ to sample images I_j. The likelihood $p(v_i|l_i)$ is computed as,

$$p(v_i|l_i) \propto \exp\left[\frac{-\gamma}{|\mathcal{I}_i|-1}\sum_{j=1}^{|\mathcal{I}_i|-1} d(h_{R(\mathbf{x}_{i,j})}, h_{R(\mathbf{x}_{i,j+1})})\right] \tag{6}$$

where $\gamma > 0$ and $d(h_{R(\mathbf{x}_{i,j})}, h^t_{R(\mathbf{x}_{i,j+1})})$ is the χ^2-distance between two HOGs $h_{R(\mathbf{x}_{i,j})}$ and $h_{R(\mathbf{x}_{i,j+1})}$. Note that $p(v_i|l_i)$ does not depend on m. Fig. 5 illustrates the computation of the likelihoods using multi-view RGB data.

Given the likelihoods $p(v_i|l_i)$ and prior $p(L_m)$ defined for each model m, the problem of repairing the shape S is to find the optimal L^* such that

$$\begin{aligned} L^* &= \underset{L_m\in\{0,1\}^{|V|}}{\arg\max}\left\{\max_{m\in\mathcal{M}} p(L_m|V)\right\} \\ &\propto \underset{L_m\in\{0,1\}^{|V|}}{\arg\max}\left\{\max_{m\in\mathcal{M}} p(V|L)p(L_m)\right\} \\ &= \underset{L_m\in\{0,1\}^{|V|}}{\arg\max}\left\{\max_{m\in\mathcal{M}}\left[\prod_{v_i\in V} p(v_i|l_i)\right]p(L_m)\right\} \end{aligned} \tag{7}$$

where $p(V|L_m)$ is replaced by $p(V|L)$ since $p(v_i|l_i)$ does not depend on m, and, similarly to conventional MRFs, it is assumed that $p(V|L) = \prod_{v_i\in V} p(v_i|l_i)$.

The problem defined in (7) is to find the best model $m \in \mathcal{M}$ that is used as the prior to maximise the posteriori $p(L_m|V)$ (i.e. the MAP inference). Since the MRF model can have cycles, the inference in (7) cannot be solved by using exact inference methods (e.g. [21]). In addition, a brute-force inference would be intractable due to exponential complexity. To overcome this issue, variational approach is adopted in the paper.

3. Variational Mean Field

Variational methods have shown their power as a robust approximation approach applied successfully in various computer vision tasks, e.g. human detection [35, 24, 25], object tracking [22], template matching [23]. In the context of graphical models, e.g. MRF [15, 13, 33], the core idea of the variational approach is to approximate the posteriori $p(L|V)$ by a variational distribution Q via maximising an objective function $J(Q)$ defined as,

$$\begin{aligned} J(Q) &= \log p(V) - KL(Q(L)||p(L|V)) \\ &= \log p(V) - \sum_L Q(L)\log\frac{Q(L)}{p(L|V)} \\ &= -\sum_L Q(L)\log Q(L) + \sum_L Q(L)\log p(L,V) \\ &= \mathcal{H}(Q(L)) + E_{Q(L)}\{\log p(L,V)\} \end{aligned} \tag{8}$$

where $\mathcal{H}(Q(L))$ is the entropy of $Q(L)$, $E_{Q(L)}\{\cdot\}$ is the expectation with respect to $Q(L)$.

Since KL is non-negative, $J(Q)$ is bounded by $\log p(V)$ and thus maximising $J(Q)$ is equivalent to retrieving both the desired marginal (i.e. $p(V)$) and the posteriori $Q^*(L)$. Indeed, if $Q^* = p(L|V)$, $J(Q^*)$ will reach the maximum. In this paper, we represent Q in the form of full factorisation (e.g. dropping edges in a Boltzmann graph) as,

$$Q(L) = \prod_{i=1}^{|V|} Q_i(l_i) \tag{9}$$

where $Q_i(l_i)$ is the variational distribution of l_i.

As defined in (2) and (4), $\psi_{m,i,j}(l_i, l_j)$ and $\psi_{m,i}(l_i)$ are expressed in the form of a Boltzmann distribution. Thus, we can write,

$$Q_i(l_i) = \mu_i^{l_i}(1-\mu_i)^{(1-l_i)} \tag{10}$$

where $\mu_i, i \in \{1, ..., |V|\}$ are computed via mean field equations [15] as follows,

$$\mu_i = \frac{p(v_i|l_i=1)k_i}{p(v_i|l_i=0) + p(v_i|l_i=1)k_i} \tag{11}$$

where $p(v_i|l_i)$ is defined in (6) and

$$k_i = \exp\left\{ \sum_{v_j \in \mathcal{N}(v_i)} \alpha f(D_m(i), D_m(j))\mu_j + \beta g(D_m(i)) \right\} \tag{12}$$

where $\mathcal{N}(v_i)$ is the set of neighbouring voxels of v_i.

As shown in (11) and (12), μ_i is updated locally based on the neighbouring nodes in $\mathcal{N}(v_i)$ and the update is performed iteratively to increase $J(Q)$ which is finally computed as,

$$\begin{aligned} J(Q) = &\sum_i \mathcal{H}(Q_i) + \sum_{i,j} \alpha f(D_m(i), D_m(j))\mu_i\mu_j \\ &+ \sum_i \beta g(D_m(i))\mu_i + \sum_i (1-\mu_i)\log p(v_i|l_i=0) \\ &+ \sum_i \mu_i \log p(v_i|l_i=1) - \log Z \end{aligned} \tag{13}$$

where $\mathcal{H}(Q_i)$ is the entropy of the individual variational distribution Q_i and $\mathcal{H}(Q) = \sum_i \mathcal{H}(Q_i)$ due to the full factorisation of Q.

The estimation of $J(Q)$, as shown in (13), requires the computation of Z, which again takes an exponential complexity. However, the optimisation of $J(Q)$ can be done without involving Z by using an alternative objective function $\widetilde{J}(Q) = J(Q) + \log Z$. Once the optimal variational distribution Q^* has been obtained, it can be used to approximate $p(L|V)$. In particular, since Q is fully factorised, we can approximate

$$p(L^*|V) \approx \prod_{i=1}^{|V|} Q_i(l_i^*) \tag{14}$$

where $l_i^* = \arg\max_{l_i} Q_i(l_i)$.

Applying (14) on the models m, the optimal configurations L_m^* with respect to m can be determined. The final configuration L^* in (7) is then selected as L_m^* which achieves the maximum of $p(L_m^*|V)$ over all models m.

4. Experiments

4.1. Implementation Details

We adopted the CDBN in [36] for learning and extracting the shape models $m \in \mathcal{M}$ as follows. The CDBN represents a shape in a $24 \times 24 \times 24$ volume with 3 voxel-pad for every dimension (resulting in a $30 \times 30 \times 30$ grid). The CDBN sequentially consists of three convolutional layers, one fully connected layer and one final layer (with 4000 hidden units as a combination of Bernoulli variables). In the network, each convolution filter is connected to all features returned by the previous layer. The CDBN was trained on the ModelNet dataset [36] including 3D CAD models from various sources such as 3D warehouse, SUN database [37], etc. Training the CDBN was conducted in a layer-wise fashion and refined using a fine-tuning method. Readers are referred to [36] for the details of the network architecture and training procedure.

Given the well-trained CDBN, a test shape is fed into the network. Voxels that belong to the object surface are set to 1 (i.e. observed voxels in the CDBN), those which are empty are set to -1 (i.e. unknown voxels in the CDBN). The CDBN then results in a set of labels using Gibbs sampling; labels 1 for foreground voxels and 0 for free space voxels. Note that the labelling is performed using only the geometric information learnt from the 3D CAD models while the appearance information from the RGB data is not taken into account. We initialise the sampling with 9 different random configurations of labels and obtain 9 labelling results. Those results are considered as the shape models (i.e. $|\mathcal{M}| = 9$). Fig. 3 shows some results of the CDBN applied on the table in Fig. 2. Note that the results may not capture the true shape of the table since only geometric information is used.

The 3D Distance Transforms are then applied on the results of the CDBN to compute the potentials $\psi_{m,i,j}(l_i, l_j)$ and $\psi_{m,i}$. In our implementation, we set $\alpha = 10$ and $\epsilon = 0.7$ in (2), $\beta = 30$ in (4). To compute the likelihoods $p(v_i|l_i)$ in (6), we sample the RGB frames so that the angle between two consecutive frames is about 30° (with a deviation of 5°). In addition, the HOGs are extracted on image regions of size 33×33 in relative to a 640×480 frame captured at about 1.2 metres from the object. This information is computed from the camera pose information. Each image region is divided uniformly into 4 sub-regions on which the HOGs are extracted. Those HOGs are then concatenated. Similarly to [3], we quantise the oriented gradients into 9 bins and compute the HOGs (of 9 bins) for 4 sub-regions to form a 36-dimensional HOG for an image region. We set $\gamma = 10$ in (6). We have experimented those parameters with various values and found that the performance was not sensitive to the changes while these settings gave good performance.

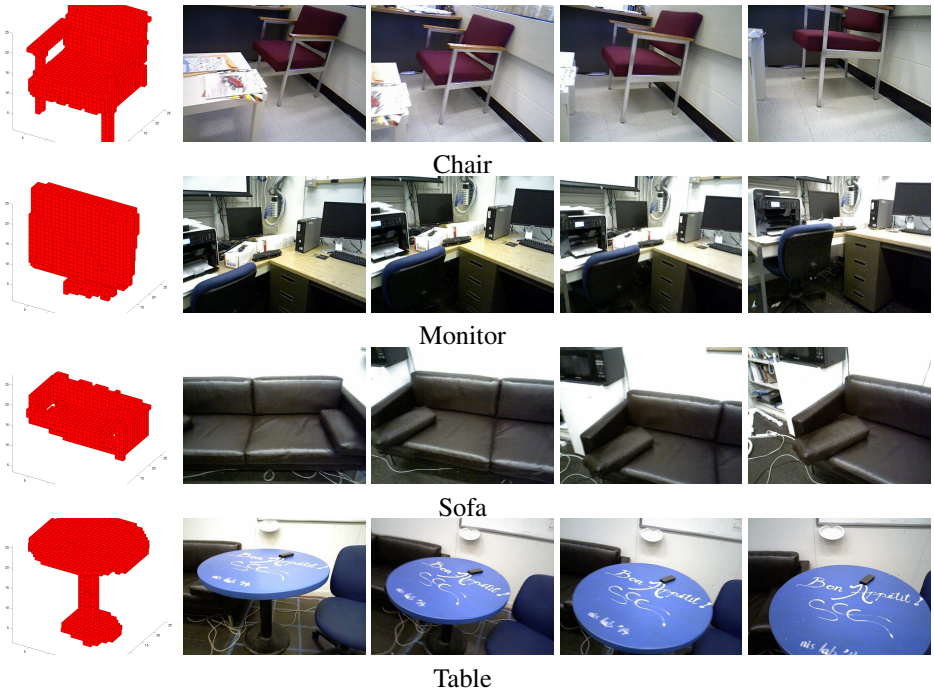

Figure 6. Some samples of our dataset: Complete 3D reconstructed model (left) and RGB images captured at different viewpoints.

For the variational mean field method, we set the maximum number of iterations to 100. However, we have observed that in our practice that the mean field approximation method could complete the inference in less then 10 iterations.

4.2. Evaluation

To evaluate the recovery ability of our proposed method, we benchmark a new 3D object dataset including 77 complete objects. Those objects are captured and reconstructed from indoor scenes. Each object is associated with a set of RGB images used to reconstruct the object. Camera poses are computed and the correspondences between 3D points and 2D pixels are also established. Fig. 6 shows some examples of complete objects in our collected dataset.

Each complete object is then degraded by randomly removing the 3D points. The removal is performed at 9 different levels varying from 10% to 90% of the original 3D points. In total, there are $77 \times 9 = 693$ objects created. In addition to the synthetic data, we also collect 10 incomplete objects reconstructed from realistic data. Fig. 7 illustrates several samples of our dataset.

To measure the performance of shape completion, Sung et al. [31] proposed two metrics: accuracy vs completeness. The accuracy measures the percentage that completed points can be matched with ground-truth points while the completeness measures the percentage that ground-truth points (after removed to create the synthetic data) can be recovered. In [31], a match is confirmed by thresholding the distances between completed points and ground-truth points. In this paper, we use the inaccuracy vs incompleteness as shape completion measures. However, instead of thresholding the distances, we directly use them in calculating the inaccuracy and incompleteness. In particular, the inaccuracy is the average of the distances from completed points to nearest ground-truth points. Similarly, the incompleteness is the average of the distances from ground-truth points to completed points. In contrast to the accuracy and completeness, the inaccuracy and incompleteness favour small distances. In other words, the smaller the inaccuracy/incompleteness is, the better the shape completion is. To efficiently compute the distances, the 3D Distance Transform is used. Fig. 8 shows the performance of our proposed method under varying levels of shape incompleteness. In this experiment, complete shapes are used as the ground truth while degraded shapes are considered as the inputs. As shown in Fig. 8, both the inaccuracy and incompleteness increase accordingly to the levels of shape degradation. However, while the inaccuracy gradually changes, the incompleteness shows a significant increase. Fig. 9 illustrates several completion results of our method.

For the current implementation, we experiment our method for $30 \times 30 \times 30$-voxel objects. However, the method is adaptive to any resolutions specified by 3D shape models. We could also apply tensor voting techniques, e.g. [34], to interpolate normals in higher resolutions. We consider this

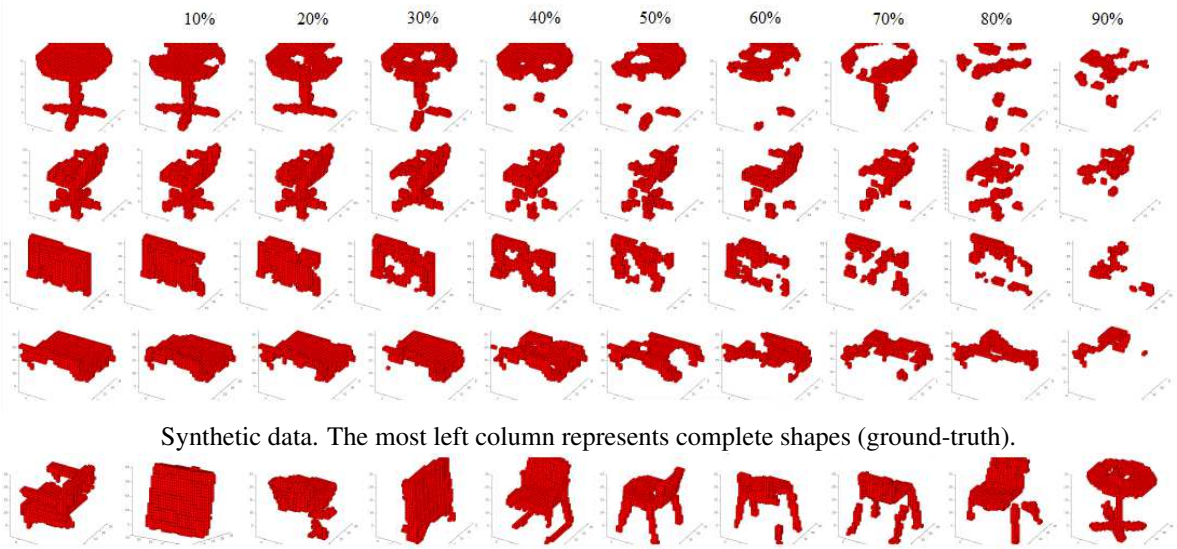

Figure 7. Some examples of incomplete 3D shapes in our dataset.

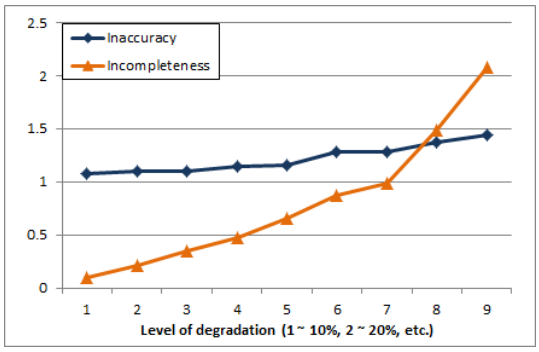

Figure 8. Performance of our method under various levels of shape degradation.

extension as our future work.

4.3. Comparison

In addition to evaluation, we also compare our method with other existing methods. In particular, we evaluate the ShapeNets [36], a recent shape completion method using high-level geometric information learnt from CAD models. For methods using low-level geometric information, we evaluate the Screened Poisson Surface Reconstruction (SPSR) [18][1] (which is shown to perform better than its original work in [17]) and the PolyMender [16][2]. Since the work in [18] makes use of gradients, in addition to the geometric information of 3D points, the normals of 3D points were also computed and buffered for use during degrading 3D objects in our dataset.

For comparison, we use the sum of both the inaccuracy and incompleteness as a single metric. Note that the ShapeNets just results in a set of 9 different shapes for a given incomplete shape. Thus, we apply our method to identify the best matching shape amongst the 9 shapes generated by the ShapeNets. The matching shape is then used for comparison. For the SPSR and PolyMender methods, completed results are voxelised to $30 \times 30 \times 30$. We report the comparison between our method and other existing methods in Fig. 10. As shown in the experiments, our method achieves the best performance. Compared with the ShapeNets purely using geometric information, our method shows the potential of multi-view RGB data in dealing with incompleteness. The proposed method also significantly outperforms the SPSR and PolyMender methods which use only low-level geometric information. This shows the benefits of the 3D shape prior in shape repairing.

4.4. Computational Analysis

We measure the complexity of the proposed method via the processing time and the number of iterations of the variational inference. Our experiments on an Intel(R) Core(TM) i7 2.10GHz CPU computer with 8.00 GB memory have shown that an incomplete shape could be repaired in about 0.03 seconds and 8.89 iterations. We have also found that both the processing time and number of iterations slightly changed under different levels of shape incomplete-

[1]The SPRS is available in MeshLab and its implementation can be found at `http://www.cs.jhu.edu/~misha/Code/PoissonRecon/Version8.0/`

[2]Binary code is available at `http://www.cse.wustl.edu/~taoju/code/polymender.htm`

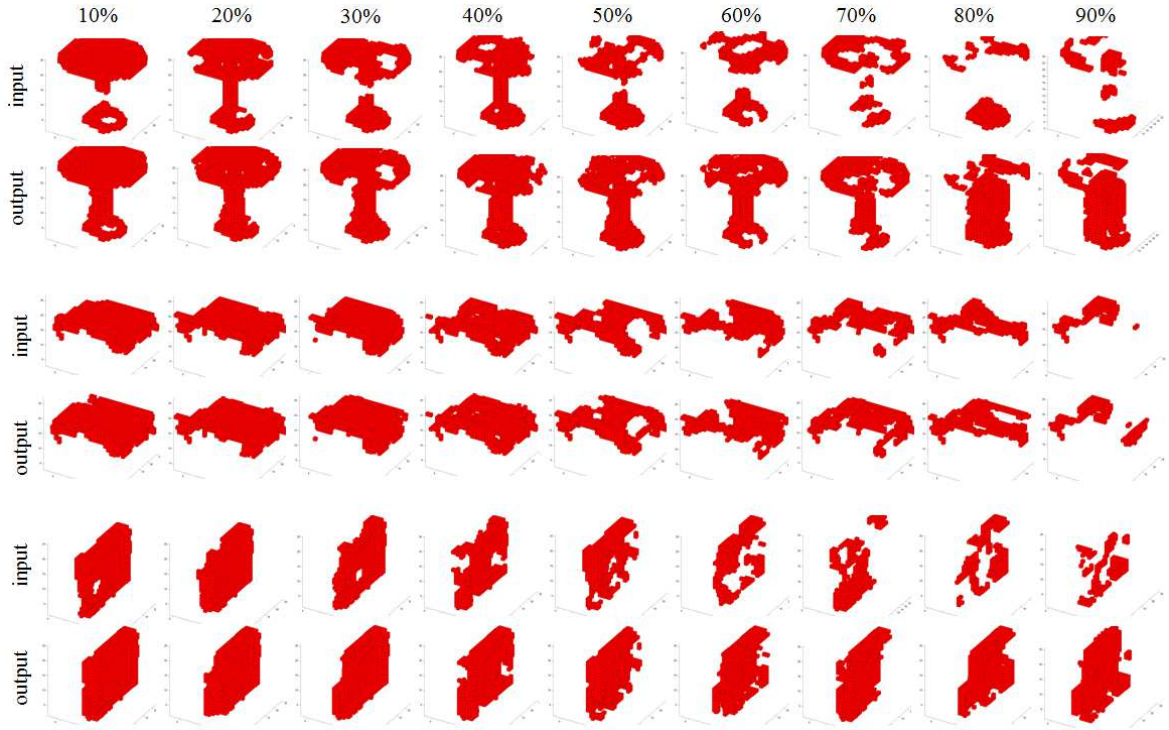

Figure 9. Some completion results of our proposed method.

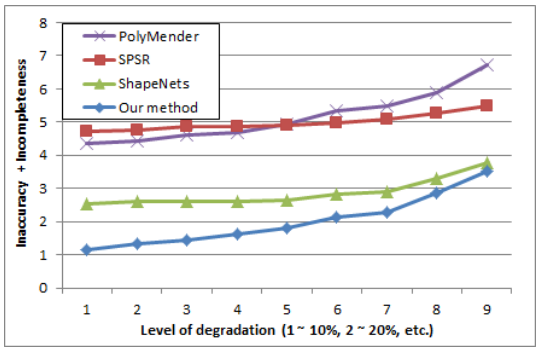

Figure 10. Comparison between our method and existing methods.

ness. In fact, those quantities depend on the result of the shape models generated by the ShapeNets.

5. Conclusion

This paper proposes a method for repairing 3D shapes using both the geometry and multi-view RGB data. The 3D shape is modelled in an MRF in which the priors between hidden nodes are obtained from shape models learnt using a convolutional deep belief network. The consistency of the RGB images of the 3D shape at multiple viewpoints is exploited in the data likelihoods. The problem of repairing an incomplete shape is formulated as the maximum a posteriori (MAP) estimation in the MRF model. Variational mean field method is used to approximation the MAP estimation. We benchmark a new 3D object dataset for evaluation of the method. Experimental results on the new dataset have shown the robustness and efficiency of the proposed method. Repairing shapes in higher resolutions would be our future work.

6. Acknowledgement

Sai-Kit Yeung is supported by Singapore MOE Academic Research Fund MOE2013-T2-1-159 and SUTD-MIT International Design Center Grant IDG31300106. We acknowledge the support of the SUTD Digital Manufacturing and Design (DManD) Centre which is supported by the Singapore National Research Foundation (NRF). This research is also supported by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative.

References

[1] X. Bai, Q. Li, L. J. Latecki, W. Liu, and Z. Tu. Shape band: A deformable object detection approach. In *CVPR*, pages 1335–1342, 2009.

[2] S. A. Barker and P. J. W. Rayner. Unsupervised image segmentation using Markov random field models. *Pattern Recognit.*, 33(4):587–602, 2000.

[3] S. Choi, Q. Y. Zhou, and V. Koltun. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[4] S. Choi, Q. Y. Zhou, and V. Koltun. Robust reconstruction of indoor scenes. In *CVPR*, pages 5556–5565, 2015.

[5] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, pages 303–312, 1996.

[6] A. Dame, V. A. Prisacariu, C. Y. Ren, and I. Reid. Dense reconstruction using 3d object shape priors. In *CVPR*, pages 1288–1295, 2013.

[7] J. Davis, S. R. Marschner, M. Garr, and M. Levoy. Filling holes in complex surfaces using volumetric diffusion. In *Symposium on 3D Data Processing Visualization and Transmission*, pages 428–438, 2002.

[8] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[9] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. Technical report, Cornell Computing and Information Science, 2004.

[10] R. Gal, A. Shamir, T. Hassner, M. Pauly, and D. Cohen-Or. Surface reconstruction using local shape priors. In *Eurographics Symposium on Geometry Processing*, pages 253–262, 2007.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

[12] C. Hane, C. Zach, A. Cohen, R. Angst, and M. Pollefeys. Joint 3D scene reconstruction and class segmentation. In *CVPR*, pages 97–104, 2013.

[13] T. S. Jaakkola. Tutorial on variational approximation methods. Technical report, MIT Artificial Intelligence Laboratory, 2000.

[14] M. Jancosek and T. Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR*, pages 3121–3128, 2011.

[15] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, pages 183–233, 1999.

[16] T. Ju. Robust repair of polygonal models. *ACM Trans. Graph.*, 23(3):888–895, 2004.

[17] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Symposium on Geometry Processing*, pages 61–70, 2006.

[18] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Trans. Graph.*, 32(3):1–13, 2013.

[19] V. G. Kim, S. Chaudhuri, L. Guibas, and T. Funkhouser. Shape2pose: Human-centric shape analysis. *ACM Trans. Graph.*, 33(4):120:1–120:12, 2014.

[20] V. G. Kim, W. Li, N. J. Mitra, S. Chaudhuri, S. DiVerdi, and T. Funkhouser. Learning part-based templates from large collections of 3d shapes. *ACM Trans. Graph.*, 32(4):70:1–70:12, 2000.

[21] F. R. Kschischang, B. J. Frey, and H. A. Loelinger. Factor graphs and the sum-product algorithm. *IEEE Trans. Info. Theory*, 47(2):498–519, 2001.

[22] C. Medrano, J. E. Herrero, J. Martínez, and C. Orrite. Mean field approach for tracking similar objects. *CVIU*, 113:907–920, 2009.

[23] D. T. Nguyen. A novel chamfer template matching method using variational mean field. In *CVPR*, pages 2425–2432, 2014.

[24] D. T. Nguyen, W. Li, and P. Ogunbona. Inter-occlusion reasoning for human detection based on variational mean field. *Neurocomputing*, 110:51–61, 2013.

[25] D. T. Nguyen, M. K. Tran, and S. K. Yeung. An mrf-poselets model for detecting highly articulated humans. In *ICCV*, pages 1967–1975, 2015.

[26] M. Pauly, N. J. Mitra, J. Giesen, M. H. Gross, and L. J. Guibas. Example-based 3d scan completion. In *Symposium on Geometry Processing*, pages 23–32, 2005.

[27] V. A. Prisacariu, A. V. Segal, and I. Reid. Simultaneous monocular 2d segmentation, 3d pose recovery and 3d reconstruction. In *ACCV*, pages 593–606, 2012.

[28] H. Roth and M. Vona. Moving volume kinectfusion. In *BMVC*, pages 1–11, 2012.

[29] A. Sharf, M. Alexa, and D. Cohen-Or. Context-based surface completion. *ACM Trans. Graph.*, 23(3):878–887, 2004.

[30] S. Song and J. Xiao. Sliding shapes for 3D object detection in depth images. In *ECCV*, pages 634–651, 2014.

[31] M. Sung, V. G. Kim, R. Angst, and L. Guibas. Data-driven structural priors for shape completion. *ACM Transactions on Graphics*, 2015.

[32] J. P. C. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, and P. H. S. Torr. Mesh based semantic modelling for indoor and outdoor scenes. In *CVPR*, pages 2067–2074, 2013.

[33] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[34] T. P. Wu, S. K. Yeung, J. Jia, C. K. Tang, and G. Medioni. A closed-form solution to tensor voting: theory and applications. *PAMI*, 34(8):1482–1495, 2012.

[35] Y. Wu and T. Yu. A field model for human detection and tracking. *PAMI*, 28(5):753–765, 2006.

[36] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1–9, 2015.

[37] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.

[38] J. Xiao, A. Owens, and A. Torralba. SUN3D: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, pages 1625–1632, 2013.

[39] Q. Y. Zhou and V. Koltun. Dense scene reconstruction with points of interest. *ACM Trans. Graph.*, 32(4):112:1–112:8, 2013.

[40] Q. Y. Zhou and V. Koltun. Simultaneous localization and calibration: Self-calibration of consumer depth cameras. In *CVPR*, pages 454–460, 2014.