A Two-Level Generative Model for Cloth Representation and Shape from Shading

Feng Han and Song-Chun Zhu
Departments of Computer Science and Statistics
University of California, Los Angeles
Los Angeles, CA 90095
hanf@cs.ucla.edu, sczhu@stat.ucla.edu

Abstract

In this paper we present a two-level generative model for representing the images and surface depth maps of drapery and clothes. The upper level consists of a number of folds which will generate the high contrast (ridge) areas with a dictionary of shading primitives (for 2D images) and fold primitives (for 3D depth maps). These primitives are represented in parametric forms and are learned in a supervised learning phase using 3D surfaces of clothes acquired through photometric stereo. The lower level consists of the remaining flat areas which fill between the folds with a smoothness prior (Markov random field). We show that the classical ill-posed problem – shape from shading (SFS) can be much improved by this two-level model for its reduced dimensionality and incorporation of middle-level visual knowledge, i.e. the dictionary of primitives. Given an input image, we first infer the folds and compute a sketch graph using a sketch pursuit algorithm as in the primal sketch [10], [11]. The 3D folds are estimated by parameter fitting using the fold dictionary and they form the "skeleton" of the drapery/cloth surfaces. Then the lower level is computed by conventional SFS method using the fold areas as boundary conditions. The two levels interact at the final stage by optimizing a joint Bayesian posterior probability on the depth map. We show a number of experiments which demonstrate more robust results in comparison with state-of-the-art work. In a broader scope, our representation can be viewed as a two-level inhomogeneous MRF model which is applicable to general shape-from-X problems. Our study is an attempt to revisit Marr's idea [23] of computing the $2\frac{1}{2}$D sketch from primal sketch. In a companion paper [2], we study shape from stereo based on a similar two-level generative sketch representation.

This is a preprint of the Statistics Department June, 2005. This version was submitted to IEEE Trans. on PAMI in 2005. A short version appeared in the IEEE Conf. on Computer Vision and Pattern Recognition, San Diego, June 2005.

I. Introduction and motivation

In this paper we present a two-level generative model for studying two classical computer vision problems: (i) shape from shading (SFS), and (ii) drapery and cloth representation. Our study has three objectives. The first is to obtain a parsimonious representation for drapery and clothes. We adopt a notion created long ago by artists [25] who paint drapery and clothes by sparse sketches and a few categories of fold primitives. These folds, in general, correspond to the ridges in the computer vision literature [14], [19], [12], [13], and they form a sketch graph as the skeleton of the drapery/clothes. The remaining areas are flat with almost no perceivable structures and therefore can be filled-in between the folds with conventional SFS method. Thus our two-level model consists of an upper level sketch graph for the folds and a lower level for the flat pixels. Without the upper level, this model reduces to the conventional SFS representation, i.e. Markov random field on pixels. Our second goal is to learn a dictionary of 3D "*fold primitives*" for surface depth maps. Each fold primitive is an elongated segment of the fold with parametric attributes for its geometric shape and surface profile (cross section) perpendicular to its axis. We learn these primitives using a number of 3D cloth/drapery surfaces acquired through photometric stereo technique. The 3D fold primitives derive a dictionary of 2D "*shading primitives*" with lighting and albedo attributes. Each shading primitive will be a rectangular image patch with relatively high shading contrast. These primitives are inferred from a single input image through parameter fitting. Our third goal is to overcome the traditional ill-posed problems associated with shape-from-shading using the lower dimensional representation and prior knowledge learned about the folds. We demonstrate robust reconstruction results (in Figures 11 and 12) in comparison with state-of-the-art SFS methods [20], [33] (in Figure 13).

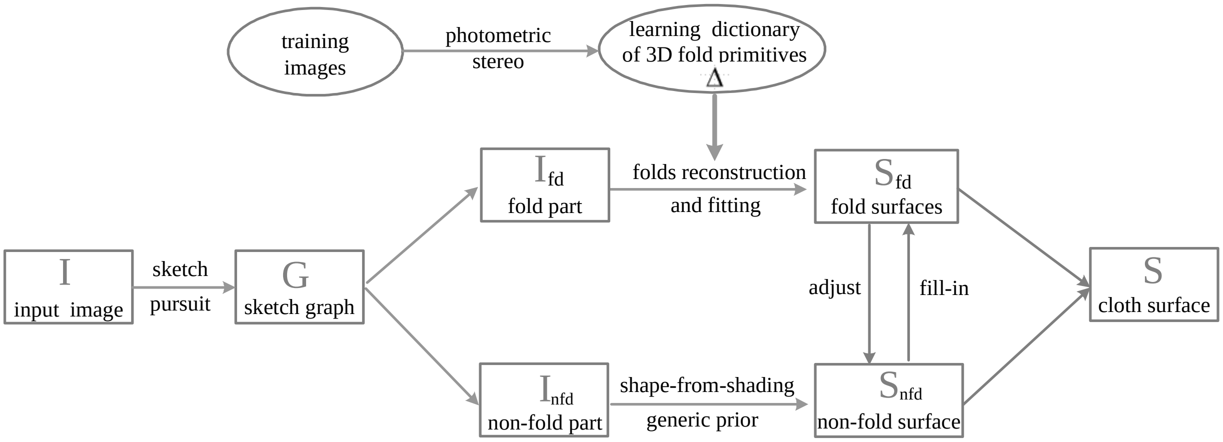


Fig. 1. The dataflow of our method for computing the 3D surface **S** of drapery/cloth from a single image **I** using the two-layer generative model. See text for interpretation.

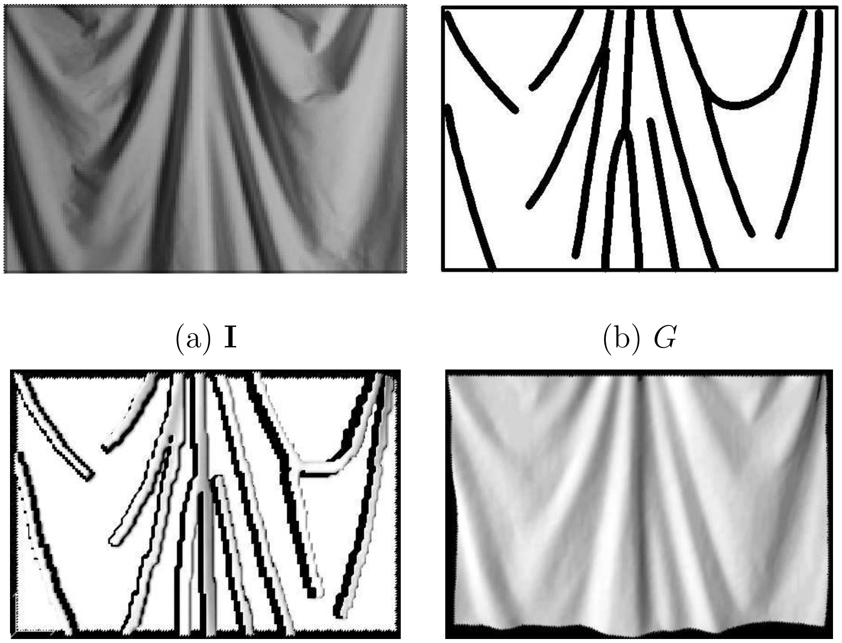
Our model is inspired by the success of a recent primal sketch model [10], [11] which represents generic images with structural and textural parts. The structural part corresponds to edges and boundaries with an attribute sketch graph which generates images with image primitives, and the textural part is modeled by a Markov random field using the structural part as boundary conditions. The two-level generative model in this paper is intended to improve the one level smoothness priors (Markov random fields) used widely in regularizing the shape-from-X tasks [15]. Thus our study is aimed at re-visiting a research line advocated by Marr [23] for computing the $2\frac{1}{2}$D sketches (surface depth maps) from a primal sketch (a symbolic token image representation with primitives). In a companion paper [2], a similar two-level generative model is studied for shape-from-stereo with a different dictionary of primitives. In the following of this section, we briefly introduce the generative model, and compare it with previous work in the literature.

A. Overview of the two-layer generative model

The dataflow of our method is illustrated in Fig. 1 and a running example is shown in Fig. 2. The problem is formulated in a Bayesian framework, and we adopt a stepwise greedy algorithm by minimizing various energy terms sequentially. Given an input image $\mathbf{I}$ on a lattice Λ, we first compute a sketch graph G for the folds by a greedy sketch pursuit algorithm. Fig. 2.(b) is an exemplary graph G. The graph G has attributes for the shading and fold primitives. G decomposes the image domain into two disjoint parts: the fold part $\mathbf{I}_{\text{fd}}$ for pixels along the sketch and non-fold part $\mathbf{I}_{\text{nfd}}$ for the remaining flat areas. Fig. 3 illustrates the two level decomposition. We estimate the 3D surface $\hat{\mathbf{S}}_{\text{fd}}$ for the fold part by fitting the 3D fold primitives in a fold dictionary Δ_{fd}. Fig.2.(c) shows an example of $\mathbf{S}_{\text{fd}}$. This will yield gradient maps $(p_{\text{fd}}, q_{\text{fd}})$ for the fold surface. Then we compute the gradient maps $(p_{\text{nfd}}, q_{\text{nfd}})$ for the non-fold part by the traditional shape-from-shading method on the lower level pixels, using gradient maps in the fold area as boundary conditions. Then we compute the joint surface $\mathbf{S} = (\mathbf{S}_{\text{fd}}, \mathbf{S}_{\text{nfd}})$ from the gradient maps (p, q) of both fold part and non-fold part. Therefore the computation of the upper level fold surfaces $\mathbf{S}_{\text{fd}}$ and the lower level flat surface $\mathbf{S}_{\text{nfd}}$ is coupled. Intuitively, the folds provide the global "skeleton" and therefore boundary conditions for non-fold areas, and the non-fold areas propagate information to infer the relative depth of the folds and to achieve a seamless surface $\mathbf{S}$. The two-level generative model reduces to the traditional smoothness MRF model when the graph G is null.

The fold dictionary is learned in an off-line training phase using a number of 3D drapery/clothes surfaces acquired by photometric stereo (see Fig.5). We manually sketch the various types of folds on the 3D surfaces (see Fig.6). These fold primitives are defined in parametric form. These 3D fold primitives derive a dictionary of *2D shading primitives* [12] for cloth images (see Fig.7) under different illumination directions.

There are three basic assumptions in the SFS literature: (i) object surfaces having con-



(a) $\mathbf{I}$ (b) G

(c) $\mathbf{S}_{\text{fd}}$ (d) $\mathbf{S}$

Fig. 2. (a). A drapery image under approximately parallel light. (b). The sketch graph for the computed folds. (c). The reconstructed surface for the folds. (d) The drapery surface after filling in the non-fold part. It is viewed at a slightly different angle and lighting direction.

stant albedo with Lambertian reflectance, (ii) single light source far from the scene, and (iii) orthographic camera projection. These assumptions are approximately observed in the drapery/clothes images in our study due to the uniform color and small relative surface depth compared with distance to the camera, except that some valleys between two deep folds have inter-reflectance and shadows. For such dark areas, we do not apply the SFS equation, but interpolate them by surface smoothness prior between the 3D folds.

B. Previous work in shape from shading and cloth representation

Shape from shading is a classical ill-posed vision problem studied in Horn's seminar work [15], [16], [18]. Most algorithms in the literature impose smoothness priors for regularization. The state of the art algorithms are given in Zheng and Chellappa [33] and others [20],

[21], [28], [31], and a survey is given in [32]. These prior models are Markov random field on pixels, which, in our opinion, are rather weak for reconstructing surfaces, and therefore more knowledge about the objects in the scene need to be incorporated for robust computation. This motivates our two-level generative model with dictionary of fold primitives. Our attempt of learning fold primitives is related to two recent work. The first is the so-called shape-from-recognition by Nandy and Ben-Arie [24] and Atick et al [1]. [1] built a probabilistic 3D model for the whole human face in an off-line learning phase, while [24] divided human face into generic parts like nose, lips and eyes and built the 3D prior models of these parts. These models are much better informed than the smoothness priors in SFS and enable them to infer 3D face surface from a single image by fitting a small number of parameters. Their work demonstrates that one can solve SFS better when high level knowledge about the object class is available. The second example is the learning of textons including "lightons" in Zhu et al [36]. This work learns generic dictionary of surface patches under varying illuminations. In comparison, our fold primitives are focused on drapery and clothes images, and we believe the same method can be extended to other object classes. The two-level model is also closely related to the primal sketch representation in Guo et al [10], [11].

Drapery/cloth representation is also an interesting problem studied in both vision and graphics, as clothes are complex visual patterns with flexible shapes and shading variations. In the graphics literature, clothes are always represented by meshes with a huge number of polygons in geometric based, physical based, and particle based cloth modeling and simulation techniques [17], [4]. Such representation is infeasible for vision purposes. In contrast, artists have long noticed that realistic drapery/clothes can be painted by a few types of folds [25]. In the vision literature, methods for detecting folds have been studied in Huggins et al [19] and Haddon and Forsyth [12], [13], mostly using discriminative method.

Besides computing the 3D surfaces, the fold representation is also useful for some applications, such as, non-photorealistic human portrait and cartoon sketch, clothed human understanding and tracking.

C. Paper Organization

The rest of the paper is organized as follows. Section II briefly reviews the formulation of shape from shading and photometric stereo. Section III presents the two-level representation for both 2D cloth images and 3D cloth surfaces. Section IV and Section V discuss the learning and inference algorithms for cloth sketching and reconstruction. Section VI shows the experiments with comparison results. Section VII concludes the paper with a discussion of limitation and future work.

II. Problem formulation of shape from shading and photometric stereo

This section discusses the formulation of two problems which will be used as components in our generative models and algorithms. (i) Shape from shading for the non-fold part. (ii) Photometric stereo used in constructing 3D cloth surface for learning the fold primitive dictionary.

A. Formulation of shape from shading

Considering a surface observed under orthographic projection, we define the coordinate system so that its x and y axes span the image plane Λ and z axis coincides with the optical axis of the camera. Thus the surface $\mathbf{S}$ is expressed as a depth map,

$$\mathbf{S}(x,y) = z(x,y), \;\; \forall (x,y) \in \Lambda. \tag{1}$$

Following the notation of Horn [15], we denote the surface gradients at (x,y) by (p,q) with

$$p(x,y) = \frac{\partial \mathbf{S}(x,y)}{\partial x}, \;\; q(x,y) = \frac{\partial \mathbf{S}(x,y)}{\partial y}. \tag{2}$$

The unit normal of the surface is,

$$\mathbf{n} = (\frac{-p}{\sqrt{p^2+q^2+1}}, \frac{-q}{\sqrt{p^2+q^2+1}}, \frac{1}{\sqrt{p^2+q^2+1}})' \tag{3}$$

For Lambertian surface with constant composite albedo (including strength of illumination and reflectivity of the surface) η, we obtain a *reflectance map* R under parallel light $L = (l_1, l_2, l_3)'$,

$$R = \eta < \mathbf{n}, L >= \eta \frac{-pl_1 - ql_2 + l_3}{\sqrt{p^2+q^2+1}} \tag{4}$$

Eqn. 4 is the basic *image irradiance equation* in SFS. R can be written as either a function $R = R(p,q,\eta,L)$ or an image $R = R(x,y)$. The light source L is a unit vector, which can be equally represented by the lighting angle (γ,τ) for the slant and tilt. So $(l_1,l_2,l_3) = (\cos\tau\cos\gamma, \cos\tau\sin\gamma, \sin\tau)$.

In the literature, the composite albedo η and light direction (γ,τ) are three global variables that can be inferred quite reliably without computing $(p,q,\mathbf{S})$. There are two statistical approaches for estimating η and (γ,τ) by Lee and Rosenfeld [21] and Zheng and Chellapa [33]. In this paper, we adopt the method in [33]. The computation of $(p,q,\mathbf{S})$ are based on the computation of η and (γ,τ) in the rest of the paper.

Since there are infinite number of surfaces that can produce the same intensity image $\mathbf{I}$, some additional constraints are needed to regularize the problem in the form of energy functions. In the literature, one may impose the smoothness energy on the surface $\mathbf{S}$, or alternatively one may put the smoothness energy plus an integrability energy on the gradients (p, q) and derive $\mathbf{S}$ from (p, q). The latter is more convenient for computation, and is adopted in this paper.

The following is an energy on (p, q) in [8] modified by smoothness weights in [30],

$$E_1(p, q) = \lambda_{\text{int}} \sum_{(x,y)\in\Lambda} (p_y - q_x)^2 + \lambda_{\text{smo}} \sum_{(x,y)\in\Lambda} (w_1 p_x^2 + w_2 p_y^2 + w_2 q_x^2 + w_3 q_y^2). \tag{5}$$

The first term is an integrability term, the second term is an inhomogeneous smoothness term, and λ_{int} and λ_{smo} are parameters balancing the two parts[1]. w_1, w_2, and w_3 are the weights used in [30], which are chosen to be inversely proportional to the intensity gradient along the x, diagonal, and y directions respectively. To compute these weights, one normalizes the image intensity to [0, 1], and calculates $w_1(x, y) = (1 - |\mathbf{I}_x(x, y)|)^2$, $w_2(x, y) = (1 - \frac{\sqrt{2}}{2}|\mathbf{I}_x(x, y) + \mathbf{I}_y(x, y)|)^2$, and $w_3(x, y) = (1 - |\mathbf{I}_y(x, y)|)^2$.

The deterministic relation in eqns.(2) is relaxed to a soft energy between $\mathbf{S}$ and (p, q),

$$E_2(\mathbf{S}|p, q) = \sum_{(x,y)\in\Lambda} (\mathbf{S}_x - p)^2 + (\mathbf{S}_y - q)^2. \tag{6}$$

The residue between a radiance image R and input image $\mathbf{I}$ is assumed to be Gaussian noise. For shadowed areas (deep and narrow valleys between folds), the image intensities (lower than a threshold δ) no longer follow the Lambertian assumption, and thus they do not contribute to the data energy term below. We denote the shadow pixels by $\Lambda_o = \{(x, y) : \mathbf{I}(x, y) \leq \delta\}$. The third energy term is,

$$E_3(\mathbf{I}|p, q, \eta, L) = \sum_{(x,y)\in\Lambda\backslash\Lambda_o} \frac{(\mathbf{I}(x, y) - R(x, y))^2}{2\sigma^2}. \tag{7}$$

[1] As a side note, the smoothness energy function together with the parameters could be learned from data using a minimax entropy principle as it was done in [35].

R is a function of p, q, η, L.

In summary, the following are the common energy terms for shape-from-shading which will be used in the non-fold area in this paper,

$$E_{\text{nfd}}(S, p, q|\mathbf{I}) = E_1(p, q) + E_2(\mathbf{S}|p, q) + E_3(\mathbf{I}|p, q, \eta, L). \tag{8}$$

As we stated before, the composite albedo η and lighting direction (γ, τ) are estimated in an initial stage. In a Bayesian formulation, energy minimization is equivalent to maximizing a posteriori probability,

$$(\mathbf{S}, p, q)^* = \arg\max p(\mathbf{I}|p, q, \eta, L)p(\mathbf{S}|p, q)p(p, q). \tag{9}$$

The three probabilities are exponential models with energies $E_3(\mathbf{I}|p, q, \eta, L)$, $E_2(\mathbf{S}|p, q)$, and $E_1(p, q)$ respectively.

In general, S, p, q are all unknowns and have to be inferred simultaneously in iterations together with the global variables η and (γ, τ). In practice, people often compute them sequentially for computational efficiency. First we compute the gradient map (p, q) by minimizing $E_1 + E_3$, and the we construct the surface map $\mathbf{S}$ by minimizing E_2.

A well-known phenomenon in SFS is the in-out ambiguity [26] that people may experience when we view the shading images. For the drapery image in Fig. 2.(a), our preception may flip between convex and concave for some parts. Such ambiguity could be resolved by the shadows in the deep valley.

B. Formulation of photometric stereo

For a static surface $\mathbf{S}$ and fixed camera, one may obtain multiple images $\mathbf{I}_i$ under varying lighting directions $\vec{L}_i$ for $i = 1, 2, ..., m$. Then we can acquire the 3D surface $\mathbf{S}$ from these images. This is the standard photometric stereo technique [27]. We shall use photometric stereo to obtain surface depth for supervised learning of the fold primitives in later section.

We write each reflectance image R_i in a $|\Lambda| \times 1$ column vector and $\mathbf{n}$ a $|\Lambda| \times 3$ matrix, then we have a group of linear equations

$$(R_1, R_2, ..., R_m) =< \eta\mathbf{n}, (L_1, L_2, ..., L_m) > . \tag{10}$$

One can solve $\mathbf{n}$ by minimizing the sum of squared errors

$$\mathbf{n}^* = \sum_{i=1}^{m} \sum_{(x,y)} (\mathbf{I}_i(x,y) - R_i(x,y))^2. \tag{11}$$

This is solved by singular value decomposition (SVD). For surface with constant reflectance η, it is well known that the SVD solution will have an in-out ambiguity. The in-out ambiguity can be resolved by the observation of shadows.

III. Drapery/cloth representation by a two-layer generative model

In this section, we introduce the folds, the two-level generative model for both the 2D cloth images and the 3D cloth surfaces, then we formulate the SFS problem in Bayesian framework with this two-level generative model.

A. The sketch graph for fold representation

The sketch graph G in Fig. 1 is an attribute graph consisting of a number of folds $f_i, i = 1, 2, ..., N_{\text{fd}}$. Each fold f_i is a smooth curve which is divided into a chain of fold primitives. Fig. 3 shows an example of the folds (in curves) and fold primitives (in deformed rectangles).

We pool all the fold primitives of G in a set V.

$$V = \{\pi_i = (\ell_i, \theta_i^{\text{geo}}, \theta_i^{\text{pht}}, \theta_i^{\text{prf}}, \Lambda_i) : i = 1, 2, ..., N_{\text{prm}}\}. \tag{12}$$

Each fold primitive $\pi_i \in V$ covers a rectangle domain $\Lambda_i \subset \Lambda$. It is usually 15-pixel long and 9-pixel wide and has the following attributes to specify the 3D surface and image intensity within its domain Λ_i.

1. A label ℓ_i indexing the three types of 3D fold primitive in the dictionary Δ_{fd} (to be introduced next section).

2. The geometric transformation θ_i^{geo} including the center location x_i, y_i, depth z_i, tilt ψ_i, slant ϕ_i, scale (size) σ_i, and deformation (shape) vector δ_i. The latter stretches the rectangle to fit seamlessly with adjacent primitives.

3. The photometric attributes θ_i^{pht} including the light source (l_{1i}, l_{2i}, l_{3i}), and surface albedo η_i. Here we assume all the fold primitives share the same light source and surface albedo. But this assumption could be relaxed in case of multiple light sources in future study.

4. The surface (depth) of each primitive is represented by the profile (cross-section) perpendicular to the principal axis of the rectangle of the primitive, and therefore is specified by parameters θ_i^{prf}. As the surface profile is represented by PCA introduced later, θ_i^{prf} includes a few coefficients for the PCA.

5. The image domain Λ_i which is controlled fully by the geometric transforms θ_i^{geo}.

B. *Two-layer generative model*

As Fig. 3.(b) shows, the graph G divides the image lattice into two disjoint parts – the fold and non-fold areas,

$$\Lambda = \Lambda_{\text{fd}} \cup \Lambda_{\text{nfd}}, \quad \Lambda_{\text{fd}} \cap \Lambda_{\text{nfd}} = \emptyset. \tag{13}$$

Thus both the image and the surface are divided into two parts,

$$\mathbf{I} = (\mathbf{I}_{\text{fd}}, \mathbf{I}_{\text{nfd}}), \ \mathbf{S} = (\mathbf{S}_{\text{fd}}, \mathbf{S}_{\text{nfd}}). \tag{14}$$

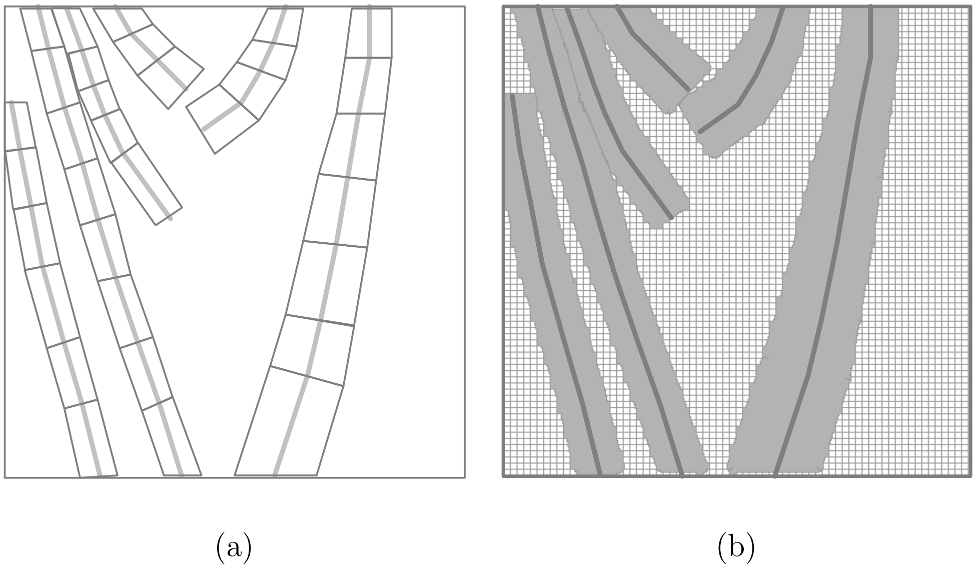


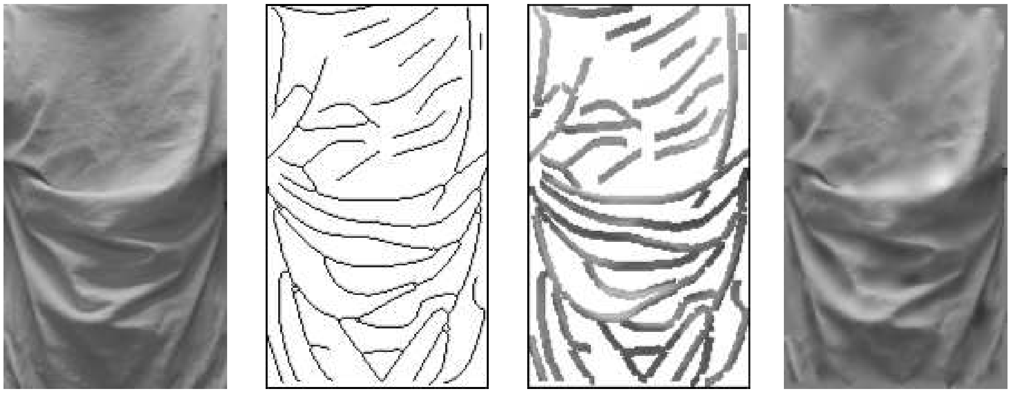
Fig. 3. (a) The graph G with rectangular fold primitives, and (b) the partition of the image domain Λ in the two-level representation of an image shown in second column of Fig. 11 (a).

In general, $\mathbf{I}_{\text{fd}}$ captures the most information and $\mathbf{I}_{\text{nfd}}$ has only secondary information and can be approximately reconstructed from $\mathbf{I}_{\text{fd}}$. Fig. 4 is an example illustrating this observation. Fig. 4.(a) is an image of clothes, (b) is the sketch graph G, and (c) is the reconstructed $\mathbf{I}_{\text{fd}}$ using shading primitives in Δ_{sd} (to be introduced soon) with attributes in G, and (d) is the reconstructed image after filling-in the non-fold part by simple heat diffusion. The heat diffusion has fixed boundary condition $\mathbf{I}_{\text{fd}}$ and therefore it does not converge to uniform image.

$$\begin{cases} \frac{\partial \mathbf{I}(x,y,t)}{\partial t} = \alpha(\frac{\partial^2 \mathbf{I}}{\partial x^2} + \frac{\partial^2 \mathbf{I}}{\partial^2 y}), & (x,y) \in \Lambda_{\text{nfd}}, \\ \mathbf{I}(x,y,t) = \mathbf{I}_{\text{fd}}(x,y), & (x,y) \in \Lambda_{\text{fd}}, \forall t \geq 0. \end{cases} \tag{15}$$

As we know, the heat diffusion is a variational equation minimizing a smoothness energy (prior). Thus filling-in the non-fold part with the probabilistic models (formulated in eqn.(9)) should achieve similar visual effects.

In the literature, it has long been observed, since Marr's primal sketch concept [23] that edge map plus gradients at edges (sketches here) contains almost sufficient image information



(a) Input (b)Folds graph G (c)$\mathbf{I}_{\text{fd}}$ (d) Filling result

Fig. 4. Fill-in the non-fold part by smooth interpolation through diffusion. This figure demonstrates the observation that the folds contains most information about the image. (a)Input cloth image. (b) The sketch graph G. (c) The fold part $\mathbf{I}_{\text{fd}}$ reconstructed from the shading primitives with G. (d) Reconstructed cloth image by filling in $\mathbf{I}_{\text{nfd}}$ by heat diffusion equation with $\mathbf{I}_{\text{fd}}$ being the fixed boundary condition.

[9] and this has been used in inpainting methods [6]. A more rigorous mathematical model for integrating the structural (sketch) with textures is referred to the recent primal sketch representation in [10], [11].

The fold area Λ_{fd} is further decomposed into a number of primitive domains

$$\Lambda_{\text{fd}} = \cup_{i=1}^{N_{\text{prm}}} \Lambda_i. \tag{16}$$

Within each primitive domain Λ_i, the surface is generated by the attributes on geometric transforms θ^{geo} and profile parameters θ^{prf},

$$\mathbf{S}_i(x,y) = \mathbf{B}_{\ell_i}(x,y;\, \theta_i^{\text{prf}}, \theta_i^{\text{geo}}), \quad (x,y) \in \Lambda_i, \tag{17}$$

where $\mathbf{B}_{\ell_i} \in \Delta_{\text{fd}}$ is a 3D fold primitive in the dictionary Δ_{fd} (next section) indexed by the type ℓ_i. $\mathbf{S}_i$ yields the gradients (p,q) and generates the radiance image using the photometric attributes $\theta_i^{\text{pht}} = (\eta_i, l_{1i}, l_{2i}, l_{3i})$,

$$R_i(x,y) = \eta_i \frac{-pl_{1i} - ql_{2i} + l_{3i}}{\sqrt{p^2+q^2+1}}, \quad (x,y) \in \Lambda_i. \tag{18}$$

In fact, we can rewrite the radiance image in terms of the shading primitive,

$$R_i(x,y) = \mathbf{b}_{\ell_i}(x,y;\, \theta_i^{\text{prf}}, \theta_i^{\text{geo}}, \theta_i^{\text{pht}}), \quad (x,y) \in \Lambda_i, \tag{19}$$

with $\mathbf{b}_{\ell_i} \in \Delta_{\text{sd}}$ being a 2D image base in the shading primitive dictionary Δ_{sd} (next section). The overall radiance image for the fold part is the mosaic of these radiance patches. In short, we have a generative model for R_{fd} on Λ_{fd} with the dictionary being parameters of the generative model.

$$R_{\text{fd}} = R(G;\Delta); \quad \Delta = (\Delta_{\text{fd}}, \Delta_{\text{sd}}). \tag{20}$$

Since each domain Λ_i often covers over 100 pixels, the above model has much lower dimensions than the pixel-based representation in conventional SFS methods. The generative relation is summarized in the following,

$$G \to \mathbf{S}_{\text{fd}} \to (p_{\text{fd}}, q_{\text{fd}}) \to R_{\text{fd}} \to \mathbf{I}_{\text{fd}}.$$

Therefore we can write the energy function for the fold part in the following,

$$E_{\text{fd}}(G|\mathbf{I}_{\text{fd}}) = E_4(\mathbf{I}_{\text{fd}}|G) + E_5(G). \tag{21}$$

E_4 is a likelihood term,

$$E_4(\mathbf{I}_{\text{fd}}|G) = \sum_{i=1}^{N_{\text{prm}}} \sum_{(x,y)\in\Lambda_i} \frac{(\mathbf{I}(x,y) - b_{\ell_i}(x,y;\Theta_i))^2}{2\sigma^2}, \tag{22}$$

where $\Theta_i = (\theta_i^{\text{prf}}, \theta_i^{\text{geo}}, \theta_i^{\text{pht}})$. E_5 is a prior term on the graph G. As G consists of a number of N_{fd} folds (smooth curves $f_1, ...f_{N_{\text{fd}}}$), $E_{\text{fd}}(G)$ penalizes the complexity K and each fold f_i is a Markov chain model whose energy is learned in a supervised way (next section).

$$E_5(G) = \lambda_0 K + \sum_{i=1}^{K} E_0(f_i). \tag{23}$$

We denote by $E_0(f_i)$ the coding length of each fold (see eqn.(28)).

So far, we have a generative model for the fold surface $\mathbf{S}_{\text{fd}}$ and radiance image R_{fd} on Λ_{fd}. The remaining non-fold part has flat surface $\mathbf{S}_{\text{nfd}}$ and nearly constant radiance R_{nfd}. One could use the traditional shape-from-shading method for the non-fold part conditioning on the fold part. Therefore we have the following objective function,

$$E_{\text{SFS2}}(G, \mathbf{S}, p, q|\mathbf{I}) = E_{\text{fd}}(G \,|\, \mathbf{I}_{\text{fd}}) + E_{\text{nfd}}(\mathbf{S}, p_{\text{nfd}}, q_{\text{nfd}} \,|\, \mathbf{I}_{\text{nfd}}, p_{\text{fd}}, q_{\text{fd}}) \tag{24}$$

The second energy is rewritten from eqn. (8). In summary, the overall problem is formulated as,

$$(G, \mathbf{S}, p, q)^* = \arg\min E_{\text{SFS2}}(G, \mathbf{S}, p, q|\mathbf{I}). \tag{25}$$

In the computational algorithm, we solve the variables in a stepwise manner. We first compute G by minimizing $E_{\text{fd}}(G \,|\, \mathbf{I}_{\text{fd}}) = E_5 + E_4$. Then with G we can derive the gradients $(p_{\text{fd}}, q_{\text{fd}})$ for the fold part, which will be used as boundary conditions in computing the gradients $(p_{\text{nfd}}, q_{\text{nfd}})$ at the non-fold part through minimizing $E_1 + E_3$. Finally we infer the joint surface $\mathbf{S}$ from (p, q) over Λ by minimizing E_2.

In the next section, we shall learn the 3D fold primitive dictionary Δ_{fd}, the shading primitive dictionary Δ_{sd}, and the Markov chain energy $E_{\text{fd}}(G)$ for the smoothness of folds.

IV. Learning the fold primitives and shape

This section discusses the supervised learning of the 3D fold primitive dictionary Δ_{fd} and its derived 2D shading primitive dictionary Δ_{sd}, and the Markov chain energy for the fold shape. They are part of the two-level generative model.

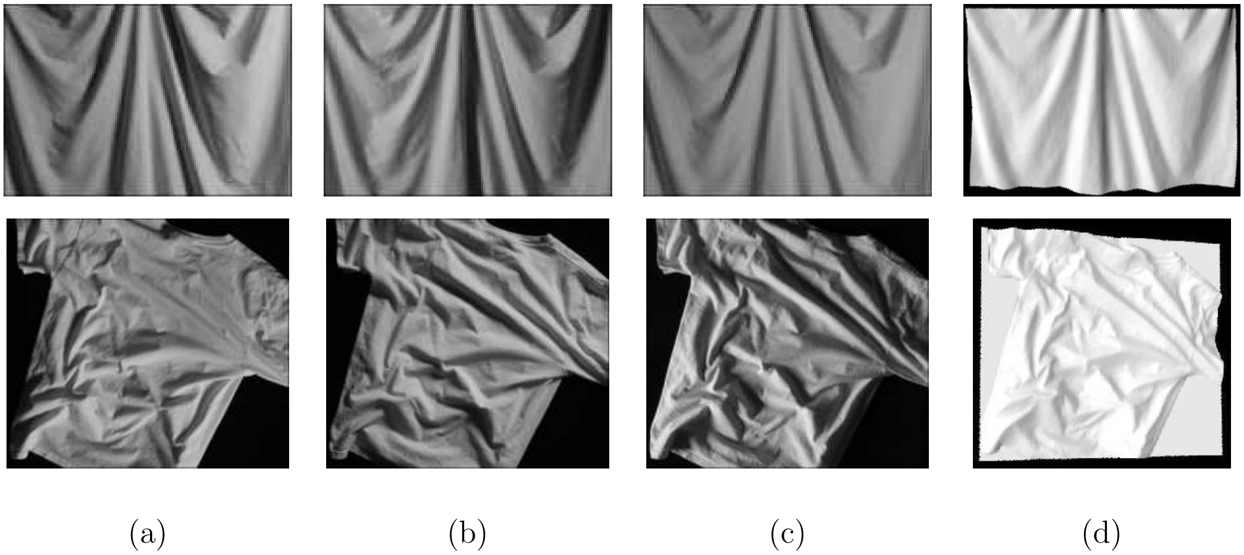


Fig. 5. (a-c) are three of the twenty images for two drapery/cloth surfaces. (d) is the reconstructed 3D surfaces using photometric stereo.

A. Learning 3D fold primitives

We use the photometric stereo algorithm [27] to acquire some 3D drapery/cloth surfaces for training. For each of drapery/cloth surface, we take about 20 images under different lighting conditions (nearly parallel light) with a fixed camera. Two examples are shown in Fig. 5 where (a-c) are three of the twenty images and (d) is the reconstructed surface. As we discussed in Section II, the photometric stereo has an in-out ambiguity for uniform reflectance surfaces. Such ambiguity can be resolved with shadow and object boundary information. For example, our perception may flip between concave and convex surfaces for the drapery (Fig.5.(a) and (b) top row).

We build an interface program to manually extract fold patches on the 3D surfaces. We draw the fold curves on the ridge (not valleys) and define the width of the fold (or scale) by rectangles whose principal axes are aligned with the ridges (See Fig. 3.(a)). We extract the surface profile at the cross-section perpendicular to the ridge (axis) within the rectangles.

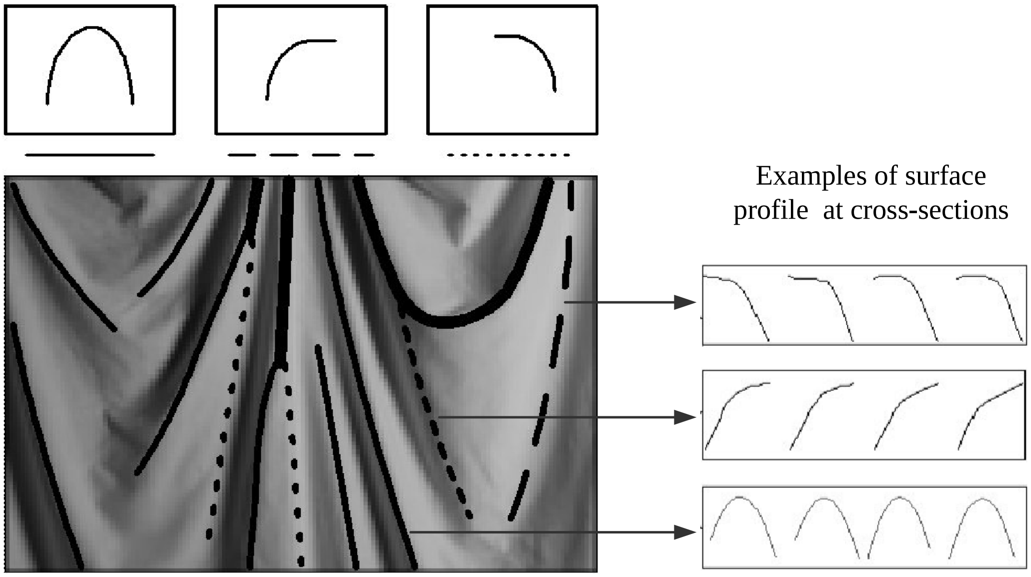


Fig. 6. Three types of fold profiles defined are shown at the top which correspond to the solid, dashed, and dotted lines respectively. The right side shows some typical examples of the surface profile at the cross-sections perpendicular to the folds.

We define three types of folds according to their surface profiles as Fig.6 shows. The first type (in solid curves) is the full fold – a ridge in frontal view. The other two types (in dashed and dotted curves) are half folds which are either a ridge in side views or the two boundaries of a broad ridge in frontal view. We show the folds on an image not the 3D surface for better view. By analogy to edge detection, the full folds are like bars and half folds are like step edges. The top row in Fig. 6 shows the surface profiles of the three types of folds. These profiles are the cross-section of the surface (from photometric stereo) perpendicular to the folds (one may call them curves, sketches, or strokes) and the right hand panel shows some examples.

Note that we only represent the ridges not the valleys for two reasons. (i) The valleys can be constructed from the ridges through surface smoothness assumption. (ii) The valleys often have heavy inter-reflectance and shadows. Their radiances are less regular than the ridges in terms of parametric modeling. We tried to introduce a fourth primitive for the shadowed deep valleys, but practically we find that such deep valleys can be reconstructed from two nearby folds.

We use principal component analysis (PCA) to represent the surface profiles. Given a number of cross-section profiles for each types of folds, we warp and interpolate to align them to have the same vector length. Then we compute the mean and a small number ($3 \sim 5$) of eigen-vectors of the profiles. So the cross-section profile is represented by a few PCA coefficients denoted by θ^{prf}. Along the fold axis, the surface is assumed to be the same (a sweep function) except that slight smoothness adjustments are needed to fit two adjacent primitives together.

Therefore, we have a dictionary of 3D fold primitives.

$$\Delta_{\text{fd}} = \{B_\ell(x, y : \theta^{\text{prf}}, \theta^{\text{geo}}) : \theta^{\text{prf}} \in \Omega_\ell, \theta^{\text{geo}} \in \Omega_{\text{geo}}, \ell = 1, 2, 3.\}. \tag{26}$$

Each fold primitive B is a 3D surface patch specified by the surface variables θ^{prf} and variables θ^{geo} for geometric transforms and deformations. In Δ_{fd}, Ω_ℓ denotes the space of the surface profiles for each type ℓ, and Ω_{geo} denotes the space of geometric transforms.

A 2D shading primitive dictionary Δ_{sd} is derived from Δ_{fd} by adding the illumination variables $\theta^{\text{pht}} \in \Omega_{\text{pht}}$.

$$\Delta_{\text{sd}} = \{b_\ell(x, y : \theta^{\text{prf}}, \theta^{\text{geo}}, \theta^{\text{pht}}) : \theta^{\text{prf}} \in \Omega_\ell, \theta^{\text{geo}} \in \Omega_{\text{geo}}, \theta^{\text{pht}} \in \Omega_{\text{pht}}, \ell = 1, 2, 3.\}. \tag{27}$$

Each shading primitive b_ℓ is an image patch generated from a surface patch $\mathbf{B}_\ell$ under lighting condition θ^{pht}. The shading primitives are similar to the "lightons" studied in [36].

As we discussed in Section (III), G is an attribute graph and each vertex is a 3D fold primitive that has all the attributes $\ell, \theta^{\text{prf}}, \theta^{\text{pht}}, \theta^{\text{geo}}$.

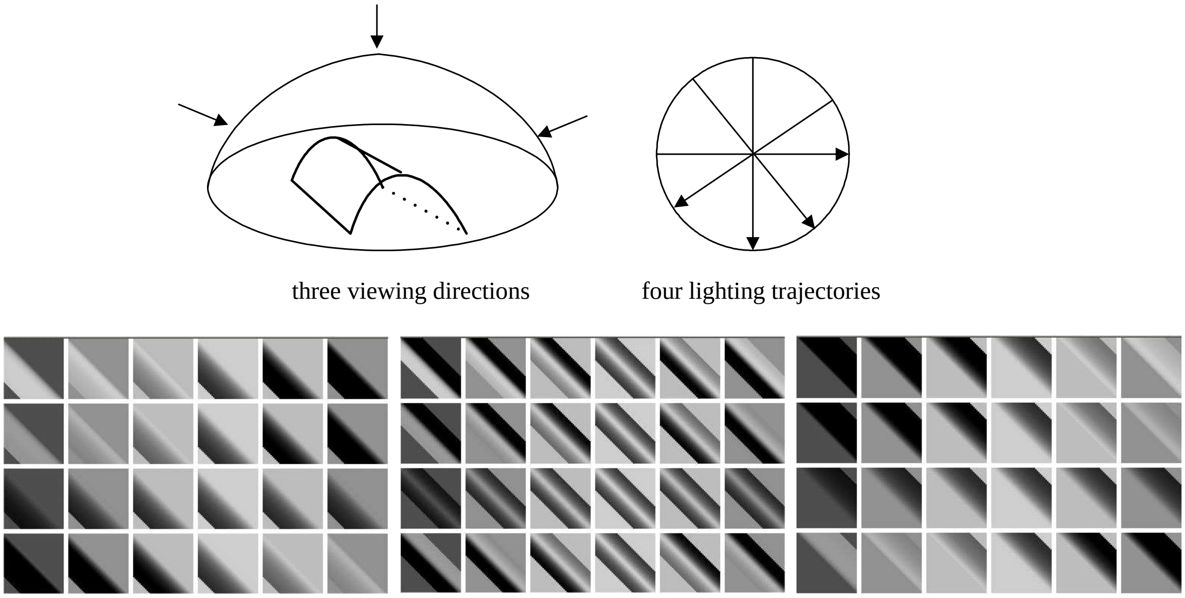


Fig. 7. The rendering results for the learned mean fold shape under different viewing directions and lighting conditions.

B. Learning shape prior for folds

The graph G consists of a number of folds $\{f_i : i = 1, 2, ..., N_{\text{fd}}\}$ with each fold f_i being a smooth attribute curve. Suppose a fold f has N_f primitives $f = (\pi_1, \pi_2, ..., \pi_{N_f})$, we model each fold as a third-order Markov chain (trigram).

$$p(f) = p(\pi_1, \pi_2) \prod_{j=3}^{N_f} p(\pi_j | \pi_{j-1}, \pi_{j-2}). \tag{28}$$

The conditional probability ensures smooth changes of the following attributes of adjacent primitives.

1. The orientation (tilt ψ and slant ϕ) in 3D space.

2. The primitives scale (width) σ.

3. The surface profile on θ^{prf}.

These probabilities are assumed to be Gaussians whose mean and covariance matrices are learned from the manually drawn folds. Then we can transform the probability $p(f)$ into energy functions E_0 in eqn. (23).

We do not model the spatial arrangements of the folds, for it leads to a three-level model as Section VII discusses (see Fig. 14).

V. Inference Algorithm for the folds and surface

The overall computing algorithm has been shown in Fig. 1. In this section, we discuss the algorithm in details. The algorithm proceeds in two phases.

In phase I, we run a greedy sketch pursuit algorithm to find all folds in the sketch graph G from the input image. At each step, the algorithm selects one shading primitive from Δ_{sd} by either creating a new fold or growing an existing fold so as to achieve the maximum reduction of the energy function following the formulation of the previous section. The energy function favors a primitive which has good fit to the dictionary and is aligned well with an existing primitive. This procedure is similar to the matching pursuit algorithm [22] in signal decomposition with wavelet dictionary, as well as the sketch pursuit in the primal sketch work [10], [11]. To expedite the pursuit procedure, we use a discriminative method [12] to detect the fold candidates in a bottom-up step. These fold candidates are proposed to the sketch pursuit algorithm which sorts the candidate primitives according to some weights and selects the primitives sequentially using a top-down generative model. This is much more effective than simply trying all shading primitives at all locations, scales and orientations in the dictionary Δ_{fd}. Therefore the algorithm bears resemblance to the data-driven Markov chain Monte Carlo methods [34], [29] except that our method is a greedy one for simplicity of the fold shape. In phase I, we also compute the 3D folds and surface $\mathbf{S}_{\text{fd}}$ by fitting the shading primitives to the 3D fold dictionary with the estimated composite albedo η and lighting direction (τ, γ).

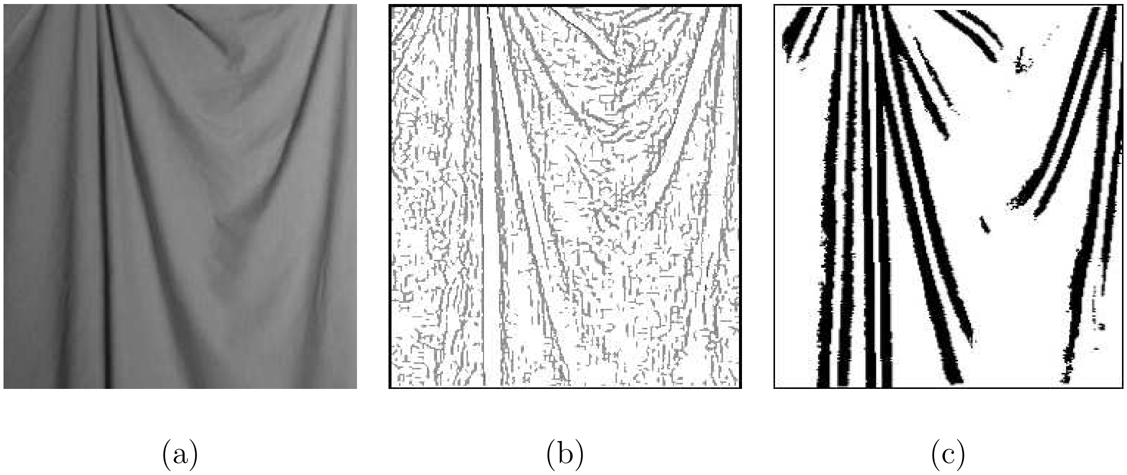


Fig. 8. (a) Input image. (b) Ridge detection result. (c) Proposed positions of folds.

In phase II, we compute the gradient map (p, q) for the non-fold areas using the existing SFS method and recover the surface $\mathbf{S}_{\text{nfd}}$ with the fold surface $\mathbf{S}_{\text{fd}}$ as boundary condition in a precess of minimizing a joint energy function.

A. Bottom-up detection of fold candidates by discriminative methods

An important feature for detecting folds is the intensity profile for the cross sections of the folds. It is ineffective to examine all possible cross sections at all possible locations, orientations, and scales. Therefore we adopt the idea of detection by a cascade of features in the recent pattern recognition literature and thus detect the fold candidates in two steps as Fig. 8 shows.

In the first step, we run a ridge detection method [14] to locate the possible locations of the ridges and estimate their local orientations. A similar method was also used in [19]. Fig. 8.(b) is an example of the ridge detection results. Then the intensity profiles at the strong ridge locations are extracted perpendicular to the ridge direction.

In the second step, we train a support vector machine [12] for binary classification – fold vs non-fold by supervised learning using the images we collected in the photometric stereo stage. Then we test the intensity profiles at various scales (widths) for the locations and orientations extracted in the ridge detection step. Fig. 8.(c) shows an example of the detected folds in the bottom-up step. As we can see in Fig. 8.(c) that the bottom-up methods detect the salient locations which will largely narrow the top-down search scope. There are two problems with the bottom-up detection: (i) disconnected segments which have to be removed or connected, and (ii) redundant points along the folds which have to be suppressed.

At each detected fold position, we do parameter fitting using the fold dictionary and get one candidate fold primitive. We denote the set of all obtained candidate fold primitives as

$$\Phi = \{(\pi_i^o, \omega_i) : i = 1, 2, ..., N_{\text{can}}\}. \tag{29}$$

Each π_i^o is a candidate fold primitive having its own label, geometric transformation and surface profile. The associated ω_i is a weight for measuring the prominence of the fold candidate which will be computed from the Bayesian probability formulation as discussed next.

B. Computing folds by sketch pursuit with the generative model

Our goal in this section is to infer the folds in the sketch graph G following the generative model in Section (III). To simplify the computation, we make two assumptions which will be released in the next step. (i) We assume the non-fold area is a flat surface $\mathbf{S}_{\text{nfd}}$ and thus has a constant radiance image $R_{\text{nfd}} = R^o_{\text{nfd}}$. $\mathbf{S}_{\text{nfd}}$ will be inferred in the next subsection, (ii) We assume the fold surface $\mathbf{S}_{\text{fd}}$ is deterministically decided by the attribute graph G, and $\mathbf{S}_{\text{fd}}$ will be adjusted in the next subsection to achieve global consistence.

Therefore, the Bayesian formulation in eqn.(25) is simplified as

$$G^* = \arg\max p(\mathbf{I}|G)p(G) = \arg\min \mathcal{L}(\mathbf{I}|G) + \mathcal{L}(G) \tag{30}$$

$\mathcal{L}(G)$ and $\mathcal{L}(\mathbf{I}|G)$ are the energy for the prior terms and likelihood respectively defined in Section (III). Here we use the notation $\mathcal{L}$ to replace E due to the new composition of energy terms.

$$\begin{aligned}
\mathcal{L}(G) &= \lambda_o K + \sum_{i=1}^{K} E_o(f_i) \\
\mathcal{L}(\mathbf{I}|G) &= \sum_{i=1}^{N_{\text{prm}}} \sum_{(x,y)\in\Lambda_i\setminus\Lambda_o} \frac{(\mathbf{I}(x,y) - b_{\ell_i}(x,y;\Theta_i))^2}{2\sigma^2} - \sum_{\mathbf{I}(x,y)\in\Lambda_{\text{nfd}}\setminus\Lambda_o} \frac{(\mathbf{I}(x,y) - R^o_{\text{nfd}}(x,y))^2}{2\sigma^2}.
\end{aligned}$$

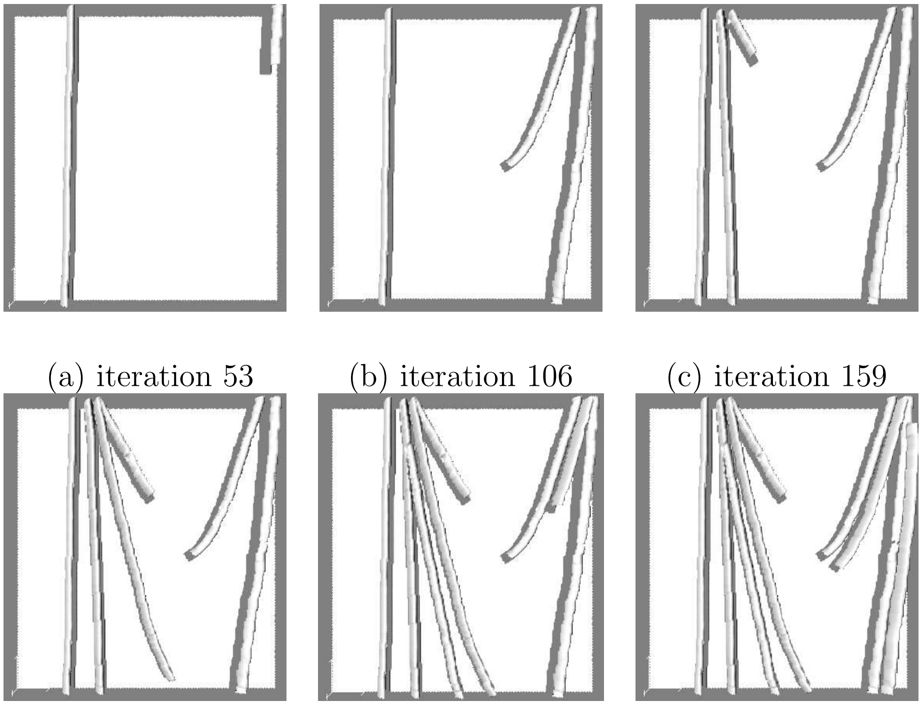
We initialize $G = \emptyset$ with the number of folds and primitives being zero ($K = N_{\text{prm}} = 0$) and $\Lambda_{\text{nfd}} = \Lambda$. Then the algorithm computes G sequentially by adding one primitive in the candidate set $\pi_+ \in \Phi$ at a time. π_+ may be a seed for starting a new fold or it may grow an existing fold f_i. Therefore we have $G_+ = G \cup \{\pi_+\}$. The primitive π_+ is selected from the candidate set Φ so that it has the maximum weight,

$$\pi_+ = \arg\max_{\pi\in\Phi}\{\omega(\pi) = \mathcal{L}(\mathbf{I}|G) - \mathcal{L}(\mathbf{I}|G_+) + \mathcal{L}(G) - \mathcal{L}(G_+)\} \tag{31}$$

The image likelihood term favors a primitive that has a better fit to a domain Λ_i against a constant image R^o_{nfd} as the background (null hypothesis). The prior term favors a primitive that has a good fit in position, scale, and orientation to existing fold according to the fold prior in eqn (28). The seed primitive that starts a new fold will receive an extra penalty λ_o in the prior term.

The weight $\omega(\pi)$ for each $\pi \in \Phi$ is initialized by the difference of the likelihood term plus a prior penalty for π being a seed for new fold. That is,

$$\omega(\pi_i) = \lambda_o + \sum_{(x,y)\in\Lambda_i\setminus\Lambda_o} \{\frac{(\mathbf{I}(x,y) - R^o_{\text{nfd}}(x,y))^2}{2\sigma^2} - \frac{(\mathbf{I}(x,y) - b_{\ell_i}(x,y;\Theta_i))^2}{2\sigma^2}\}. \tag{32}$$



(a) iteration 53 (b) iteration 106 (c) iteration 159

(d) iteration 212 (e) iteration 265 (f) iteration 317

Fig. 9. The state of sketch graph G at iteration 53, 106, 159, 212, 265, 317 on the cloth image shown in Figure 8.

Each time after a new primitive π_+ (with domain Λ_+) is added, the weights of other primitive candidates $\pi_i \in \Phi$ will remain unchanged unless in the following two cases.

Case I: π_i overlaps with π_+. As the pixels in $\Lambda_i \cap \Lambda_+$ has been explained by π_+, then for π_i the likelihood term in eqn. (32) only sums over $\Lambda_i \setminus \Lambda_+$. Therefore the weight $\omega(\pi_i)$ usually is reduced after adding π_+ to G.

Case II: A few neighboring candidate primitives which fit well with π_+ and can be considered the next growing spot from π_+. For such neighbors π_i, the prior term penalty λ_o is then replaced by the less costly smoothness energy in the Markov chain model $E_o(f)$ in eqn. (28). Therefore the weight $\omega(\pi_i)$ is increased after adding π_+ to G

The candidates with changed weights are sorted according to the updated weights, and

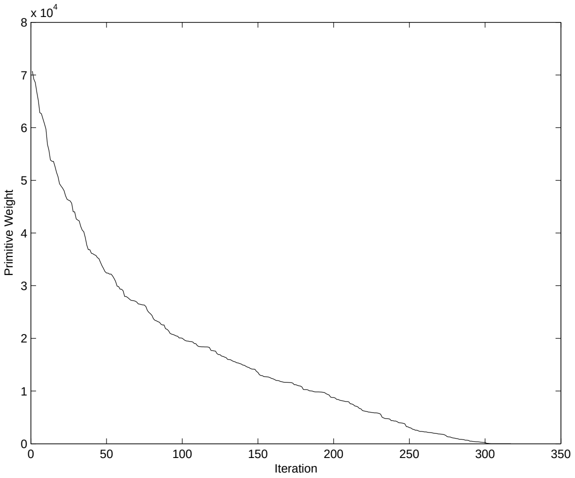


Fig. 10. The plot of the weight $\omega(\pi_+)$ (vertical axis) for the newly added primitive π_+ at each iteration (horizontal axis).

the procedure stops when the heaviest candidate has weight less than a threshold $\omega(\pi_+) \leq \delta_o$ ($\delta_o = 0$ in our experiment). Figure 9 shows an experiment for the sketch pursuit procedure at 53, 106, 159, 212, 265, and 317 iterations respectively. The plot of the weight $\omega(\pi_+)$ at each iteration is shown in Figure 10.

To summarize, we have the following algorithm for pursuing the sketch graph G.

The algorithm for sketch pursuit process

1. Bottom-up phase I: ridge detection for fold location and orientation.
2. Bottom-up phase II: SVM classification of fold primitives to obtain a candidate set Φ.
3. Initialize $G \leftarrow \emptyset$, $\Lambda_{\text{nfd}} = \Lambda$, and $R_{\text{nfd}} = R^o_{\text{nfd}}$.
4. Add a new primitive to G. $G \leftarrow G \cup \{\pi_+\}$ with highest weight

   $\omega(\pi_+) = \arg\max\{\omega(\pi) : \pi \in \Phi\}$.
5. Update the weights $\omega(\pi)$ for the neighboring primitives $\pi \in \Phi$ of π_+.
6. Re-sort the candidate primitives according to their weights in decreasing order.
7. Repeat 4-6 until $\omega(\pi_+) \leq \delta_o$.

After computing G, we obtain the fold primitives from the shading primitives by fitting the radiance equation with the estimated composite albedo η and lighting direction (γ, τ). Therefore we obtain an estimate for the fold surface $\hat{\mathbf{S}}_{\mathrm{fd}}$. One problem with the estimated depth map $\hat{\mathbf{S}}_{\mathrm{fd}}$ is that the depth of each primitive is estimated locally in the domain Λ_i. Thus it has one degree of freedom for its absolute depth. As a result, the relative depth between any two folds f_i, f_j are undecided. The relative depth must be decided together with the non-fold areas in solving the whole surface.

C. Infer the non-fold surface and refine the fold surface

Given the estimated fold surface $\hat{\mathbf{S}}_{\mathrm{fd}}$, we estimate the gradient map $(p_{\mathrm{fd}}, q_{\mathrm{fd}})$ on Λ_{fd} which is not affected by the undecided depth constant at each fold primitive. In the next step, we infer the gradient maps $(p_{\mathrm{nfd}}, q_{\mathrm{nfd}})$ for the non-fold surface Λ_{nfd} using $(p_{\mathrm{fd}}, q_{\mathrm{fd}})$ on Λ_{fd} as boundary condition. Thus we rewrite the energy in eqn.5 in the following,

$$E_1(p_{\mathrm{nfd}}, q_{\mathrm{nfd}} | p_{\mathrm{fd}}, q_{\mathrm{fd}}) = \lambda_{\mathrm{int}} \sum_{(x,y)\in\Lambda_{\mathrm{nfd}}} (p_y - q_x)^2 + \lambda_{\mathrm{smo}} \sum_{(x,y)\in\Lambda_{\mathrm{nfd}}} (w_1 p_x^2 + w_2 p_y^2 + w_2 q_x^2 + w_3 q_y^2). \quad (33)$$

We take a gradient method in [8] for minimizing the energy with respect to $(p_{\mathrm{nfd}}, q_{\mathrm{nfd}})$. At each iteration, we compute the gradient $\nabla E_1(p_{\mathrm{nfd}}, q_{\mathrm{nfd}})$ and choose a discrete stepsize d so that the energy is minimized along the gradient direction. The procedure stops until the decrease of energy is less than a threshold proportion to the number of $|\Lambda_{\mathrm{nfd}}|$.

Now we have the gradient map (p, q) for the whole image domain Λ, and we compute the surface $\mathbf{S} = (\mathbf{S}_{\mathrm{fd}}, \mathbf{S}_{\mathrm{nfd}})$ by minimizing the following energy term rewritten from eqn.6,

$$E_2(\mathbf{S}|p, q) = \sum_{(x,y)\in\Lambda} (\mathbf{S}_x - p)^2 + (\mathbf{S}_y - q)^2. \quad (34)$$

We solve the above minimization problem using the just discussed gradient method in [8]. This will be our final result for the surface.

To conclude this section, we list the steps of the overall algorithm. The algorithm seeks for greedy steps for estimating the variables $\eta, (\gamma, \tau), G, \mathbf{S}_{\text{fd}}, (p_{\text{nfd}}, q_{\text{nfd}}), \mathbf{S}_{\text{nfd}}$ sequentially by minimizing some energy terms.

The overall algorithm for SFS with the two-level generative model

1. Estimate the global variables: the composite albedo η and lighting direction (γ, τ).
2. Compute the sketch graph G by the fold pursuit process.
3. Initialize the fold surface $\hat{\mathbf{S}}_{\text{fd}}$ by fitting the radiance equations.
4. Compute the gradient map $(p_{\text{fd}}, q_{\text{fd}})$ for the fold part.
5. Compute the gradient map $(p_{\text{nfd}}, q_{\text{nfd}})$ for the non-fold part conditioned on $(p_{\text{fd}}, q_{\text{fd}})$.
6. Estimate the cloth surface $\mathbf{S} = (\mathbf{S}_{\text{fd}}, \mathbf{S}_{\text{nfd}})$ jointly from (p, q).

It is valuable in theory to estimate the variable in iterations for more accurate result which maximizes the Bayesian posterior probability jointly. For example, adjusting the estimation of η, γ, τ or G after computing $\mathbf{S}$. In practice, we find these greedy steps often obtain satisfactory results, and therefore do not pursue the maximization of the joint posterior probability.

VI. Experiments

We test our whole algorithm on a number of images. Figure 11 shows the results for three images of drapery hung on wall and a cloth image (last column) on some people. Figure 12 shows four cloth images on people. The lighting direction and surface albedos for all the testing cloth are estimated by the method in [33].

In the experimental results, the first row are input images, the second row are the sketches of folds in the input images and their domain, the third row are the syntheses for $\mathbf{I}_{\text{fd}}$ based on the generative sketch model for the fold areas, the fourth row are the 3D reconstruction

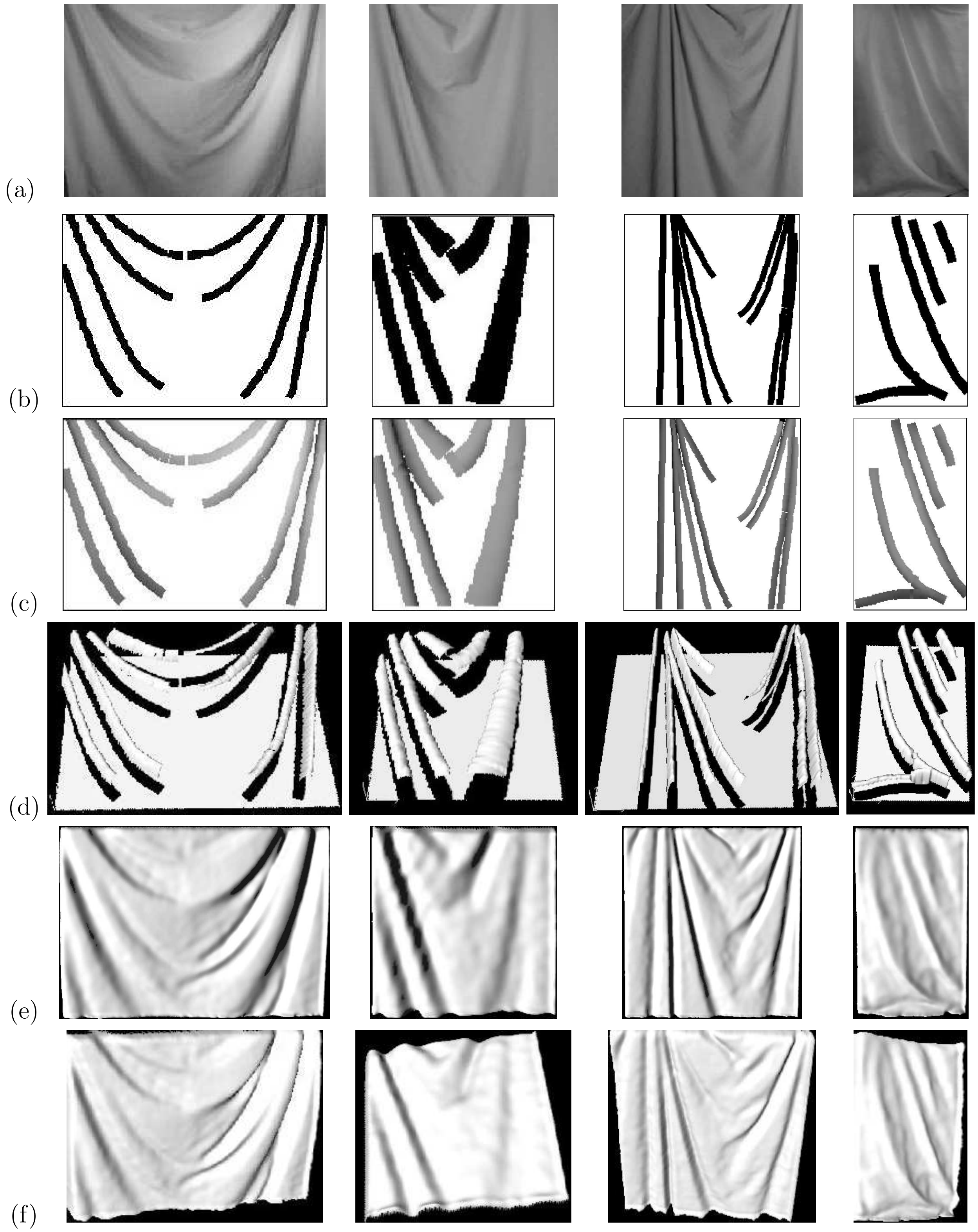


Fig. 11. (a). Input cloth image. (b). 2d folds and their image domains. (c). Synthesis for 2D fold sketches $\mathbf{I}_{\rm fd}$. (d). 3D reconstruction $\mathbf{S}_{\rm fd}$ for fold areas. (e-f). Final reconstructed surface $\mathbf{S}$ in novel views.

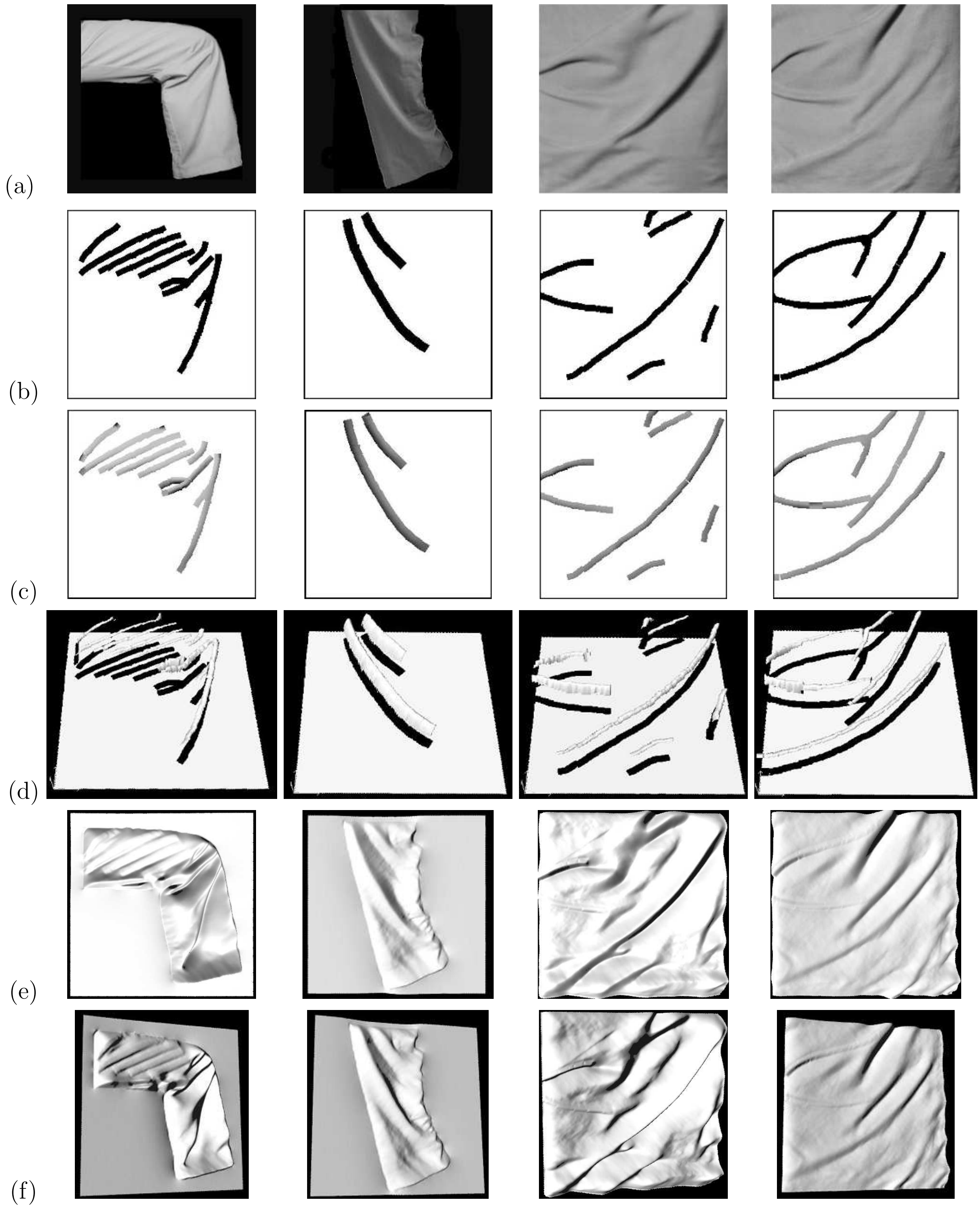


Fig. 12. (a). Input cloth image. (b). 2d folds and their image domains. (c). Synthesis for 2D fold sketches $\mathbf{I}_{\text{fd}}$. (d). 3D reconstruction $\mathbf{S}_{\text{fd}}$ for fold areas. (e-f). Final reconstructed surface $\mathbf{S}$ in novel views.

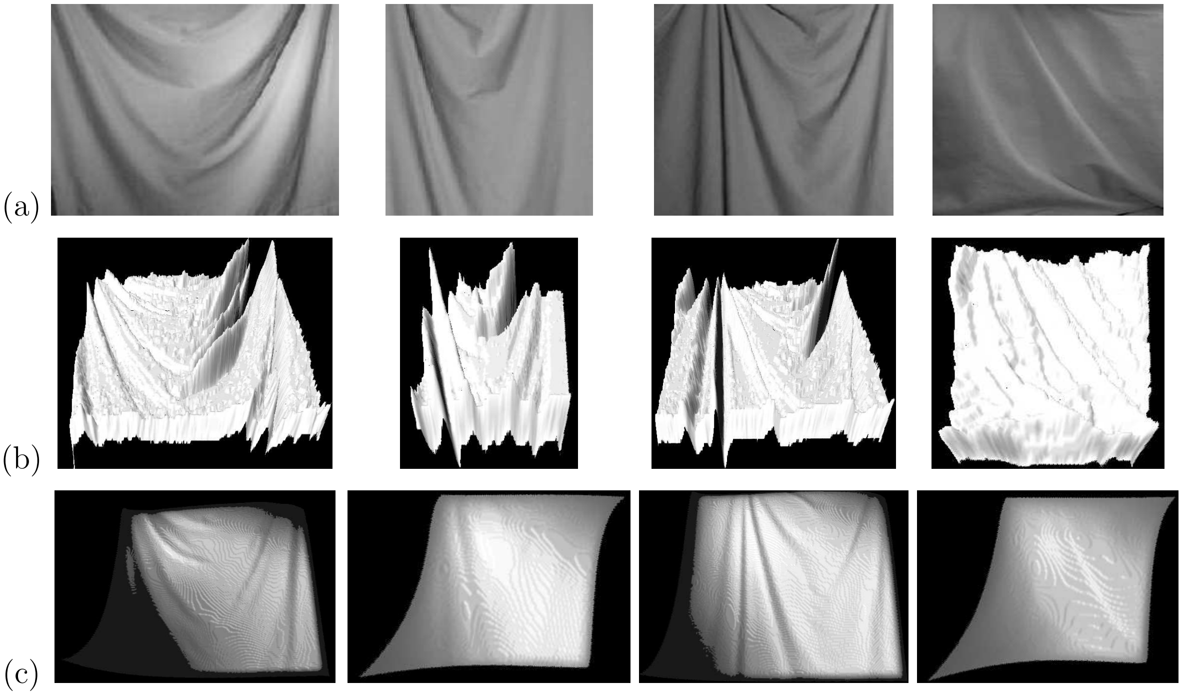


Fig. 13. (a). Input cloth images. (b). Cloth reconstruction results by approach in (Zheng and Chellappa, 1991). (c). Cloth reconstruction results by approach in (Lee and Kou, 1993).

results $\mathbf{S}_{\text{fd}}$ for the fold areas, while fifth and sixth rows are the final reconstruction results of the whole cloth surface $\mathbf{S}$ shown in two novel views.

In these results, the folds in row (d) have captured most of the perceptually salient information in the input images, and they can reconstruct the surface without too much skewing effects. It makes sense to compute them before the non-fold part. We observe that the SFS for the non-fold parts indeed provides useful information for the 3D positions of the folds.

For comparison with the state-of-art-art SFS algorithms, we run two minimization approaches on the same testing images above, as the minimization approaches are more robust and accurate even though much slower than other approaches. The first approach is from [33], while the second one is from [20]. The results for these two approaches are shown in second and third rows respectively in Figure 13. The main problems with these methods are: (i) the global surface shape could be skewed (see row (c)) due to the accumulated effect of computing $\mathbf{S}$ from the gradient map (p, q) through integration over pixels; and (ii) the relative depth of the folds are not computed reliably. This latter problem is overcome in our method due to the use of fold dictionary.

VII. Discussion and future Work

The two-level generative model is an attempt to incorporate visual knowledge representation for solving the traditional ill-posed vision problems. The current representation is still limited and shall be extended in the following directions.

1. One may infer a third level generative representation that will explain the spatial relation of the folds. As Fig. 14 shows, the folds are not independently distributed, but are radiated from some hidden structures which artists called "hubs" [25]. These hubs are shown in dashed ellipses for protruding parts or grip points such as the shoulder, or curves for boundaries such as a waist belt. A fold must start or end with one of these hubs. Physically, these hubs are the stretching points or lines that create the cloth folds.

2. We should learn a richer set of surface and shading primitives for more object classes. One extreme will be the human face model in shape-from-recognition [1], [24], and the other related work is the study of textons and lightons in [36]. With a richer dictionary, we expect that SFS could work for various locations in an image with different types of object primitives even without assuming global parallel lighting conditions.

Acknowledgements

This work was supported in part by National Science Foundation grants IIS-0222967 and IIS-0244763. We thank Jinhui Li for collecting some of the drapery and cloth images used

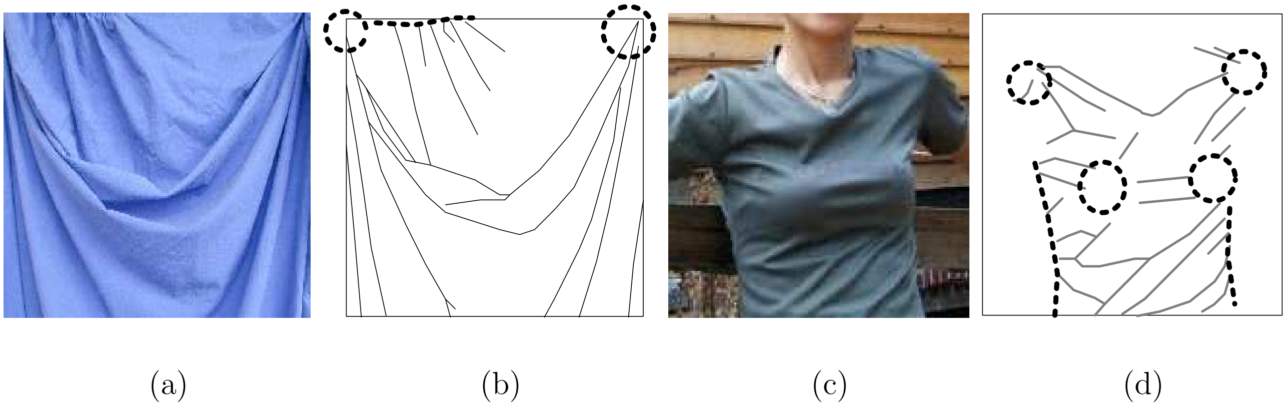


Fig. 14. (a) and (c) are respectively a drapery and a cloth image image. (b) and (d) are the folds with additional dashed ellipses and lines which represent a third hidden level for stress areas and boundaries that cause the folds for the image in (a) and (c). They are the sources and sinks for the folds. This calls for a three-level generative model.

in photometric stereo and shading, and thank Dr. Yingnian Wu for stimulating discussions.

References

[1] J.J. Atick, P.A. Griffin and A.N. Redlich "Statistical Approach to Shape from Shading: Reconstruction of 3-Dimensional Face Surfaces from Single 2-Dimensional Images", *Neural Computation*, vol. 8, no. 6, pp. 1321-1340, August 1996.

[2] A.B. Barbu and S.C. Zhu, "Incorporating Visual Knowledge Representation in Stereo Reconstruction", Submitted to ICCV, 2005.

[3] P.N. Belhumeur and D. Kriegman, "What Is the Set of Images of an Object Under All Possible Illumination Conditions?", *Int'l J. Computer Vision*, vol. 28, no. 33, pp. 24560, 1998.

[4] K. Bhat, C. D. Twigg, J. K. Hodgins, P.K. Khosla, Z. Popvic, and S.M. Seitz, "Estimating Cloth Simulation Parameters from Video", *Proc. Symposium on Computer Animation*, pp. 37-51, 2003.

[5] V. Blanz and T. Vetter, "A Morphable Model for the Synthesis of 3D Faces", *Computer Graphics Proceedings SIGGRAPH*, pp. 187-194, 1999.

[6] T. Chan and J. Shen, "Local inpainting model and TV inpainting", *SIAM J. of Applied Math.*, 62:3, 1019-43, 2001.

[7] C. Cortes and V. Vapnik, "Support Vector Networks", *Machine Learning*, vol. 20, no. 3, pp. 273-297,

1995.

[8] A. Crouzil, X. Descombes and J.-D Durou, "A Multiresolution Approach for Shape From Shading Coupling Deterministic and Stochastic Optimization", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, pp. 1416-1421, Nov. 2003.

[9] J. H. Elder, "Are edges incomplete?", *Int'l, J. of Computer Vision*, 34(2/3), 97-122, 1999.

[10] C.E. Guo, S.C. Zhu and Y.N. Wu, "Towards a Mathematical Theory of Primal Sketch and Sketchability", *Proc. Int'l Conf. Computer Vision*, pp. 1228-1235, 2003.

[11] C.E. Guo, S.C. Zhu and Y.N. Wu, "Primal Sketch: Integrating Texture and Structure", Preprint No. 416, Department of Statistics, UCLA, 2005.

[12] J. Haddon and D. Forsyth, "Shading Primitives: Finding Folds and Shallow Grooves", *Proc. Int'l Conf. on Computer Vision*, pp. 236-241, 1998.

[13] J. Haddon and D. Forsyth, "Shape Representations From Shading Primitives", *Proc. European Conf. on Computer Vision*, pp. 415-431, 1998.

[14] R. Haralick "Ridges and Valleys on Digital Images", *Computer Vision, Graphics, and Image Processing*, vol. 22, no. 10, pp. 28-38, Apr. 1983.

[15] B.K.P. Horn and M.J. Brooks *Shape from Shading*, MIT Press, 1989.

[16] B.K.P. Horn, "Height and Gradient From Shading", *Int'l J. Computer Vision*, vol. 5, no. 1, pp. 37-75, 1990.

[17] D.H. House and D.E. Breen, *Cloth Modeling and Animation*, A.K. Peters, Ltd., 2000.

[18] K. Ikeuchi, and B.K.P. Horn, "Numerical Shape From Shading and Occluding Boundaries," *Artificial Intelligence*, vol. 17, pp. 141-184, 1981.

[19] P.S. Huggins, H.F. Chen, P.N. Belhumeur, and S.W. Zucker, "Finding Folds: On the Appearance and Identification of Occlusion", *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 718-725, 2001.

[20] K.M. Lee and C.C.J. Kuo, "Shape from Shading with a Linear Triangular Element Surface Model", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 8, pp. 815-822, August 1993.

[21] C.H. Lee and A. Rosenfeld, "Improved methods of estimating shape from shading using the light source coordinate system", *Shape from Shading*, MIT Press, pp. 323-569, 1989.

[22] S. Mallat and Z. Zhang, “Matching Pursuit in a Time-Frequency Dictionary”, *IEEE Trans. on Signal Processing*, 41, 3397-3415, 1993.

[23] D. Marr, *Vision*, W.H. Freeman and Company, NY, 1982.

[24] D. Nandy and J. Ben-Arie “Shape from Recognition: A Novel Approach for 3D Face Shape Recovery”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.10, no. 2, pp. 206-217, Feburary 2001.

[25] K. Nicolaides, *The Natural Way to Draw*, page 109-117, Houghton Mifflin Com, Boston, 1941.

[26] J. Oliensis, “Uniqueness in Shape from Shading”, *Int'l J. Computer Vision*, 6(2):75-104, June, 1991.

[27] A. Shashua, *Geometry and Photometry in 3D Visual Recognition*, Ph.D Dissertation, MIT, 1992.

[28] P.S. Tsai and M. Shah, “Shape from Shading Using Linear Approximation”, *Image and Vision Computing*, vol. 12, no. 8, pp. 487-498, October 1994.

[29] Z.W. Tu and S.C. Zhu, “Image Segmentation by Data-Driven Markov Chain Monte Carlo”, *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol.24, no.5, pp. 657-673, May, 2002.

[30] G.Q. Wei and G. Hirzinger, “Parametric Shape-From-Shading by Radial Basis Functions”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 353-365, April 1997.

[31] P.L. Worthington and E.R. Hancock, “New Constraints on Data-Closeness and Needle Map Consistency for Shape-from-Shading”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 12, pp. 1250-1267, Dec. 1999.

[32] R. Zhang, P.-S. Tsai, J.E. Cryer, and M. Shah, “Shape from Shading: A Survey”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, pp. 1416-1421, November. 1999.

[33] Q. Zheng and R. Chellappa, “Estimation of Illumination Direction, Albedo, and Shape from Shading”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 7, pp. 680-702, July 1991.

[34] S. C. Zhu, R. Zhang, and Z. W. Tu, “Integrating Top-down/Bottom-up for Object Recognition by Data Driven Markov Chain Monte Carlo”, *Proc. of CVPR*, Hilton Head, SC. June, 2000.

[35] S.C. Zhu and D.B. Mumford, “Prior Learnig and Gibbs Reaction-Diffusion”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.19, no.11, pp1236-1250, Nov. 1997.

[36] S.C. Zhu, C.E. Guo, Y.Z. Wang and Z.J. Xu, “What are Textons?”, *Int'l J. Computer Vision*, vol. 62, no. 1/2, pp. 121-143, 2005.