Finding Nemo: Deformable Object Class Modelling using Curve Matching

Mukta Prasad[1] Andrew Fitzgibbon[2] Andrew Zisserman[3] Luc Van Gool[1]
[1]ETH Zürich {mprasad,vangool}@ee.ethz.ch
[2]Microsoft Research, Cambridge awf@microsoft.com
[3]University of Oxford az@robots.ox.ac.uk

Abstract

An image search for "clownfish" yields many photos of clownfish, each of a different individual of a different 3D *shape in a different pose. Yet, to the human observer, this set of images contains enough information to infer the underlying* 3D *deformable object class. Our goal is to recover such a deformable object class model directly from unordered images.*

For classes where feature-point correspondences can be found, this is a straightforward extension of non-rigid factorization, yielding a set of 3D *basis shapes to explain the* 2D *data. However, when each image is of a different object instance, surface texture is generally unique to each individual, and does not give rise to usable image point correspondences. We overcome this sparsity using curve correspondences (crease-edge silhouettes or class-specific internal texture edges).*

Even rigid contour reconstruction is difficult due to the lack of reliable correspondences. We incorporate correspondence variation into the optimization, thereby extending contour-based reconstruction techniques to deformable object modelling. The notion of correspondence is extended to include mappings between 2D *image curves and corresponding parts of the desired* 3D *object surface. Combined with class-specific priors, our method enables effective deformable class reconstruction from unordered images, despite significant occlusion and the scarcity of shared* 2D *image features.*

1. Introduction

Community photograph collections are a rich source of information about the world, particularly when many different photos of the same subject are captured over time and space. The Photosynth system [14] shows how several views of the same rigid structure (e.g. "Pantheon", "Half-dome") can be interrelated via the common coordinate system of a 3D reconstruction. We would like to extend the subject of such reconstructions beyond rigid structures to include object *classes*: sets of 3D model instances drawn from some common distribution. Such classes occur often in nature, *e.g.* the class of oak leaves, or lily petals, or dolphins. When each photo in the input collection is of a *different instance* of the object class, as might be returned by a (*e.g.* Flickr) query on the tag "oak leaf", each photo corresponds to a different 3D shape.

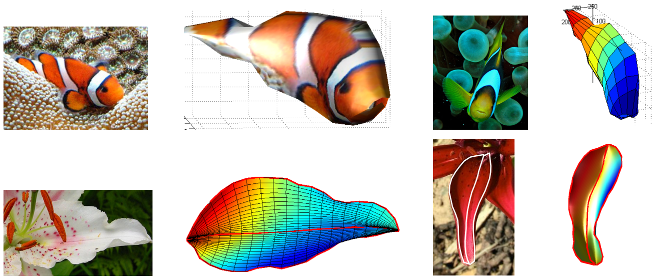


Figure 1. Deformable shape models with 4 bases for the 'lily' and 'clownfish' are recovered and fitted to some examples.

In this paper, we address the novel problem of 3D object *class* reconstruction from multiple, unordered images. Fig. 1 shows images from two classes: "lily" and "clownfish", with the 3D models recovered from each by our method. Specifically, we focus on a subset of such categories, namely objects for which a wireframe description is appropriate. In the absence of reliable point correspondences, we exploit class-specific curve correspondences. We extend existing optimization methods to incorporate variable correspondences and partial occlusion, while matching such curves. We first recapitulate existing work in § 2 before expanding on the problem in § 3.

2. Related work

Related research comprises several strands: non-rigid structure from motion, rigid 3D reconstruction from contours, and the modelling of shape classes, *e.g.* using active shape models.

2.1. Non-rigid structure from point matching

The recovery of non-rigid structure from motion has been an ongoing subject of research [4, 3, 5, 16]. Common to all these approaches, is the need for a set of point correspondences, or point tracks in the case of video. The 3D shape in each view is modeled as a linear combination of *basis shapes*, denoted $\mathcal{B}_{1..K}$. The shape in each input view is given by linear combination coefficients α, possibly with associated transformation parameters $\mathtt{R}$, $\mathbf{T}$. A set of P point correspondences over N input images is represented as a $2P \times N$ *measurement matrix*, which concatenates the 2D points $\mathbf{w}_{pn}$. Recovery of the unknown model parameters α *etc.*, was initially cast as a matrix factorization problem, but more recent work has cast it as a maximum *a-posteriori* (MAP) estimation of the parameters [5] or maximum likelihood distribution fitting [16]. Rabaud and Belongie [12] is a nice departure from the linear subspace model, but remains in a regime where features are fully in correspondence.

These techniques are very effective in analyzing a single object moving non-rigidly in video. However, applying these techniques to object class learning from community photo collections incurs a number of difficulties. First, obtaining correspondences automatically is difficult without the temporal coherence of video, or the matching tensors of rigid-body multiple-view geometry [8]. Even if manually-supplied matches are allowed, matches must be found which are consistent *across* the object class, not just per-object. The type of surface texture found on natural objects, such as floral petals and faunal pelts, tends to be unique to each object instance. This means that correspondences between different instances, as required in this paper, cannot be found. In practice, for the petal example we use throughout this paper, perhaps two reliable point correspondences may be identified: the base and tip of the petal.

Bartoli *et al.* [1] is probably the closest prior work to ours, in that they augment point-based Non-Rigid Structure from Motion (NRSfM) with curve correspondences. Their work, however, uses many more point correspondences to guide the estimation, makes strong use of the temporal constraints in video, and does not deal with curve occlusions.

2.2. Rigid structure from curve matching

In the absence of surface correspondences, then, how may the problem be approached? One solution is to move from zero-dimensional point matching to one-dimensional curve matching. The number of matched items remains low—just three to eight reliable across-class curves for our examples—but each curve correspondence provides much richer information about 3D shape than a point correspondence. The difficulty with curve correspondences, however, is a variant of the aperture problem: although the curve is in correspondence as a whole, individual points on the curve are not naturally in correspondence. Measures such as curvature provide poor matching constraints, as curvature can change drastically under projection (consider a smooth helical segment which is imaged with a cusp). Projective concepts such as bi-tangency do not extend to the deformable case.

Existing work on curve matching uses the constraints associated with rigid-body assumption to constrain the matching. Schmid and Zisserman [13] showed how the use of 2- and 3-view matching tensors allows correspondence transfer: a point on one curve which is not parallel to an epipolar line, induces a point match on the corresponding curve in a second view, and constrains the local matching homography, allowing surface texture adjacent to the curve match to resolve ambiguities. Rigid curve matching in multiple (> 3) views is addressed by Berthilsson *et al.* [2], who introduce a bundle adjustment strategy to allow curve correspondences to vary along the image curves. Kaminski and Shashua [9] derive constraints on algebraic curves from multiple views, while Martinsson *et al.* [10] combine curve fitting and reconstruction for planar curves.

3. Multiple-view reconstruction of curve families

This paper combines several of the above themes for effective recovery of deformable shape models from random collections of images. Higher-level features such as image-based curves are used and the process of joint optimization over an analytic objective (using bundle adjustment [17]) includes the search for correspondences (in addition to the basis shapes, cameras *etc.*). Instead of restricting ourselves to 2D–2D correspondences, we also incorporate 2D–3D correspondences, thus allowing for a large range of occlusion. We exploit class-specific regularization and topological information for effective reconstruction on two distinct object classes.

We first define the problem in §3.1. The general goal (§3.2) and an overview of the optimization method (§4) are described next. The specific objective, regularization and optimization change depending on the specific application. Variations of the basic method are applied to two classes in §5: "Lily" (§5.1) and "Clownfish" (§5.2). We then summarize our contributions (§6) and discuss shortcomings and future directions.

3.1. Problem statement

We have N images, each a different instance of an object class, and the goal is: (i) to extract a deformable shape model, (ii) find the camera projection parameters, and (iii) fit the shape model to each image. Each image is of a different class instance, therefore identification of point correspondences is often impossible. However, image contours of characteristic texture edges and silhouettes succinctly capture the image-based information for the task of building a deformable shape model. In each image n, a different number f_n of unique 2D curves can be easily extracted as illustrated in fig. 3 (a). Each curve is represented analytically as a piecewise-smooth, spline-based function $\boldsymbol{\omega}_{ni}(t), t \in [0,1], i \in 1 \ldots f_n$ mapping the unidimensional spline parameter t, $[0,1] \rightarrow \mathbb{R}^2$ to the image locations (fig. 2).

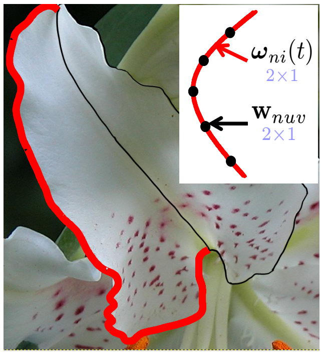

$\underset{2\times1}{\boldsymbol{\omega}_{ni}(t)}$
$\underset{2\times1}{\mathbf{w}_{nuv}}$

Figure 2. Sub-pixel edgel chains are represented by piecewise-smooth spline functions; creating continuous contours. For $u \in \{1 \ldots U\}, v \in \{1 \ldots V\}, i \in \{1 \ldots f_n\}, t \in [0,1]$.

We use a parametric 3D surface representation for the 3D shapes, *i.e.* a mapping from (u,v) parameter space to 3D, denoted $X(u,v)$. Each image-based 2D curve $\boldsymbol{\omega}_{ni}(t)$ is also a projection of some 3D curve on the object surface, which is in turn represented by a 2D curve $(u(t), v(t))$ in the surface's parameter space, which we call the *pre-image*. Provided these curves are not too close in parameter space, we may reparametrize so that the pre-images are lines. Thus there is a mapping: $t \in [0,1] \leftrightarrow$ line $[(u_{ni1}, v_{ni1}), (u_{ni2}, v_{ni2})]$, between the 2D image curve parameter and the 3D curve parameter (see fig. 3). In this work, ω_{ni} constitutes the entire, usable image information. Thus, the set $\{\boldsymbol{\omega}_{ni}(t) | i \in \{1 \ldots f_n\}\}$ and their 3D counterparts form the "visible" part of the object surface in image n. For the 2D parametric surface defined in the next section, this mapping defines the visibility of a surface vertex (p,q) in each image n as:

$$\psi_{npq} = \begin{cases} i & \text{if } (p,q) \in \text{line}\left[(u_{ni1}, u_{ni2}), (v_{ni1}, v_{ni2})\right] \\ 0 & \text{otherwise} \end{cases}$$

or, in other words: ψ_{npq} is the index of the curve in image n to which parameter-space point (p,q) maps.

Surface representation: The parametric surface representation mapping $[0,1]^2 \mapsto \mathbb{R}^3$ is used to represent fully freeform shape instances. The 3D model for the n^{th} image is represented on a discretized parametric grid as a $U \times V$ vertex mesh, $\mathbf{X}_{nuv} = [X_{nuv}, Y_{nuv}, Z_{nuv}]^\top$, $u \in \{1 \ldots U\}, v \in \{1 \ldots V\}$. (see fig. 3). The model for the 3D deformable object class follows the literature, and is a linear combination of a set $\mathcal{B}$ of K *basis shapes*; the $(u,v)^{th}$ vertex in the k^{th} basis given by: $\{\mathbf{B}_{kuv}\}_{k=1}^K$. $\mathcal{B}$ is fitted to each image by a vector of shape parameters $\boldsymbol{\alpha}_n$ to retrieve the relevant 3D model vertices given by: $\mathtt{X}_{nuv} = \sum_{k=1}^K \alpha_{nk} . \mathtt{B}_{kuv}$. We adopt the convention that $\alpha_{n1} = 1$, $\forall n$, so that $\mathtt{B}_1$ behaves like the mean in principal components analysis. The model is projected into the image via a 3×4 camera matrix: $\mathtt{P}_n = [\mathtt{A}_n \mid \mathbf{T}_n]$ and perspective projection $\pi(x,y,z) := (x/z, y/z)$ gives us the current predicted projections $\hat{\mathbf{w}}_n(u,v)$:

$$\underset{2\times1}{\hat{\mathbf{w}}_{nuv}} = \pi\left(\mathtt{A}_n \sum_{k=1}^K \alpha_{nk} \cdot \underset{3\times1}{\mathbf{B}_{kuv}} + \underset{3\times1}{\mathbf{T}_n}\right) \tag{1}$$

3.2. Desiderata

The goal is to recover the unknowns $\Theta = \{\boldsymbol{\alpha}_{1..N}, \mathtt{B}_{1..K}, \mathtt{P}_{1..N}\}$. For any solution it is important for the individual reprojections to be consistent with information from the corresponding images. For under-constrained problems appropriate regularization encourages the reconstructions $\mathtt{X}_n$ to have desired characteristics of class shape *e.g.* smoothness and topology. To this end, we will minimize a sum of reprojection error (E_{RP}) and regularization (E_{smooth}) in the combined objective

$$E_{\text{gen}} = E_{\text{RP}} + E_{\text{smooth}}. \tag{2}$$

3.3. Reprojection error

The projection $\hat{\mathtt{W}}_n$ of the n^{th} 3D model $\mathtt{X}_n$ must be consistent with its image observations. If the corresponding image projection $\mathbf{w}_{nuv}$ for each $\mathbf{X}_{nuv}$ were known, the model fit could be assessed by measuring reprojection error:

$$E_{\text{RP}} = \sum_{n,u,v} \|\underset{2\times1}{\mathbf{e}_{nuv}}\|^2 = \sum_{nuv} (\psi_{nuv} > 0) \cdot \left\| \underset{2\times1}{\hat{\mathbf{w}}_{nuv}} - \underset{2\times1}{\mathbf{w}_{nuv}} \right\|^2 \tag{3}$$

Fitting the model by minimizing E_{gen} is then a straightforward bundle adjustment over $\Theta = \{\boldsymbol{\alpha}_{1..N}, \mathtt{B}_{1..K}, \mathtt{P}_{1..N}\}$. However we lack point correspondences, so the closest-point from a projected point to a image-based curve represents the "correspondence" and the modified reprojection error is:

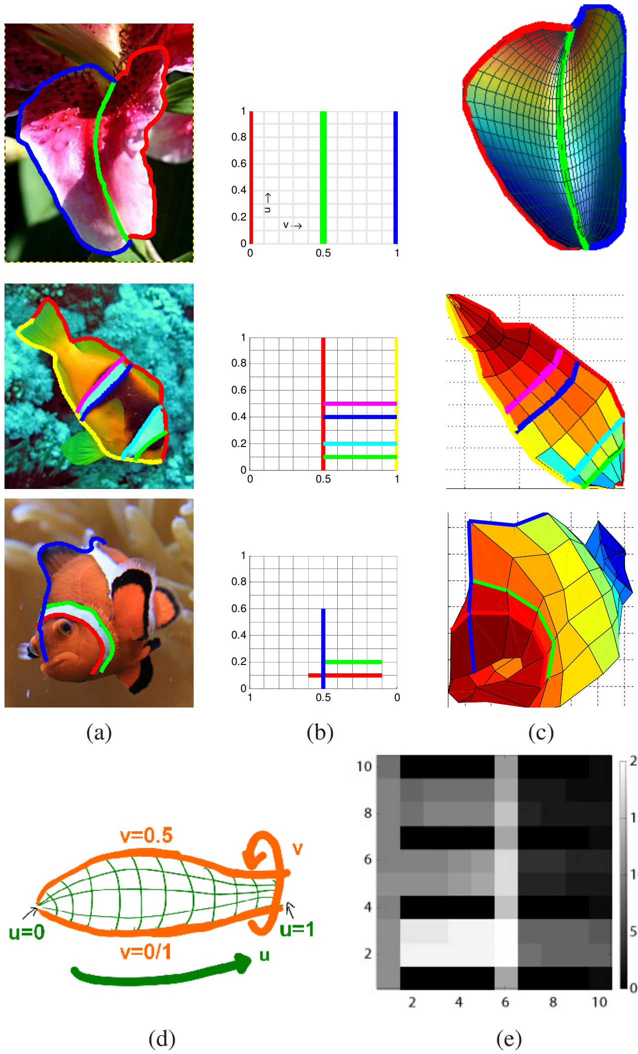
$$\mathbf{d}_{nuv}(t) = \left(\hat{\mathbf{w}}_{nuv} - \boldsymbol{\omega}_{n,\psi_{nuv}}(t)\right) . (\psi_{nuv} > 0) \tag{4}$$

$$\mathrm{D} = \sum_{n,u,v} \|\mathbf{d}_{nuv}^{\min}\|^2 \tag{5}$$

$$\mathbf{d}_{nuv}^{\min} = \min_t \|\mathbf{d}_{nuv}(t)\| \tag{6}$$

This small modification makes the optimization rather more difficult. Several options are available to minimize D, and these are the topic of the next section.

Minimizing D: In our experiments, a number of strategies for minimization of D were considered: (i) distance transform, (ii) point-to-spline distance, and (iii) augmented bundle adjustment. The distance transform approach maintains a distance look up table (and associated derivatives as



(a) (b) (c)

(d) (e)

Figure 3. **Annotation and parametrization:** (a) Image and their prominent curve features. (b) Mapping between image curves and parameter grid. For lilies, this is constant $\{(u,v)|u \in [0,1], v \in \{0, 0.5, 1\}\}$. Row 1: The gray vertices are interpolated from the ribs post-optimization. For clownfish, the curves are intuitively and approximately mapped to the cylindrical fish topology (also see fig. 2). Rows $2, 3$: Black vertices are the invisible (or occluded) vertices *w.r.t.* the image curve information of (a) and are solved simultaneously with the coloured ones. **Note:** The image curves are unique to each instance and identically-coloured curves correspond only along the columns and not across rows. (d) Cylindrical topology for fish parametrization. (e) For each vertex, the brightness corresponds to the number of images it is visible in. Only two $(u,v) = \{(2,6),(3,6)\}$ are visible $\forall N$.

in [6]): $\mathrm{DT}_{ni}(\mathbf{x}) = \min_t \|\mathbf{x} - \boldsymbol{\omega}_{ni}(t)\|$. This gives the simplification

$$\mathbf{d}^{\min}_{nuv} = (\psi_{nuv} > 0) \cdot \mathrm{DT}_{n,\psi_{nuv}}(\hat{\mathbf{w}}_{nuv}). \tag{7}$$

Discretization is a problem on distance transforms. To increase accuracy, we can discretize more finely (by a factor of 5), but this leads to memory constraints on the number of images we can train on. The advantage of this look-up based approach is that it works very fast.

Because the image curves are defined as interpolating splines, the closest point to any query point can be efficiently and reliably found using Newton-Raphson iterations. If we assume each 3D point is determined by its projections, then (6) can be explicitly minimized at each function evaluation. However, the derivatives of the function are not available in closed-form, so finite-difference approximations must be used, and bundle adjustment becomes considerably slower.

The third strategy is to augment the bundle adjustment with NUV extra parameters $t_N(u,v)$, *i.e.* rewriting

$$D_{\text{RP}} = \min_{\Theta} \sum_{nuv} \min_t d_{nuv}(t) = \min_{\Theta, t_1(u,v)..t_N(u,v)} \sum_{nuv} d_{nuv}(t). \tag{8}$$

Each $t_n(u,v)$ represents the closest curve-point to each visible $\hat{\mathbf{w}}_n(u,v)$ and adds NUV parameters (assuming all vertices are visible) to the optimization to explicitly represent the correspondences. This redundancy removes the need for explicit closest point computations. However, it does not greatly increase the computational load because the additional parameters add a large sparse block (of tightly constrained variables) to the Jacobian. In our experiments, the additional block does not add any report-worthy time to Jacobian calculation.

3.4. Regularization

In the presence of sparse training data with occlusion and noise, regularizers on the 3D shape are required. Each $\mathtt{X}_n$ (also written as the $3UV \times 1$ vector $\vec{\mathbf{X}}_n$) must be smooth regardless of which vertices are visible, and possess desirable class characteristics for plausible reconstruction. Thin-plate energies (see [15, 18, 7]) associated with first (tension) and second (bending energy) derivatives and their corresponding matrix operators: $\mathtt{C}_u, \mathtt{C}_v, \mathtt{C}_{uu}, \mathtt{C}_{uv}, \mathtt{C}_{vv}$ (each $3UV$ square matrix as shown in [11]) are defined ignoring parametrization issues:

$$E_{\text{bending}} = \sum_n \|\mathbf{e}^n_{\text{bending}}\|2 \tag{9}$$

$$\underset{9UV\times1}{\mathbf{e}^n_{\text{bending}}} = \lambda \cdot \left[\vec{\mathbf{X}}^\top_{n_{uu}} \quad \sqrt{2}\vec{\mathbf{X}}^\top_{n_{uv}} \quad \vec{\mathbf{X}}^\top_{n_{vv}}\right]^\top \tag{10}$$

$$E_{\text{tension}} = \sum_n \|\mathbf{e}^n_{\text{tension}}\|^2 \tag{11}$$

$$\underset{9UV\times1}{\mathbf{e}^n_{\text{tension}}} = \phi \cdot \left[\vec{\mathbf{X}}^\top_{n_u} \quad \vec{\mathbf{X}}^\top_{n_v}\right]^\top \tag{12}$$

$$E_{\text{smooth}} = E_{\text{bending}} + E_{\text{tension}} \tag{13}$$

$$\underset{3\times 1}{\mathbf{X}_{n_u}}(i,j) = \frac{1}{2U}\left(\mathbf{X}_{n(i+1)j} - \mathbf{X}_{n(i-1)j}\right) \tag{14}$$

$$\underset{3\times 1}{\mathbf{X}_{n_{uu}}}(i,j) = \frac{1}{4U^2}\left(\mathbf{X}_{n(i+1)j} - 2\mathbf{X}_{nij} + \mathbf{X}_{n(i-1)j}\right) \tag{15}$$

$$\text{and, } \underset{3UV\times 1}{\mathtt{X}_{n_{uu}}} = \underset{3UV\times 3UV}{\mathtt{C}_{uu}} \underset{3UV\times 1}{\vec{\mathbf{X}}_n} \; ; \; |||^{ly} \text{ for others} \tag{16}$$

4. Hierarchical Optimization

Given a dataset of images and curves, computing a solution for the deformable object problem involves an optimization over a sum-of-squares objective of type (2), which can expressed as a residual vector. The Levenberg-Marquardt algorithm is used with an analytically computed Jacobian. Fig. 4 illustrates the main sparsity structure that is exploited in our experiments. We divide the matrix into its major blocks, many of which are block-structured, and process those blocks in "tetris mode", *i.e.* efficiently compute and compactly store the blocks in column-wise order in a flattened matrix, increasing the optimization speed and accuracy. For under-constrained non-convex problems the key to a good solution is the choice of initialization. We reduce the dependency by having an hierarchical minimization strategy with incremental model complexity. Relatively reliable solutions can be found for simpler cases (*e.g.* rigid body *i.e.* $K = 1$, scaled-orthographic projection). This we relax towards the full solution by varying K from 1 to the desired target, and for each K, we first estimate the newly introduced bases and coefficients before proceeding to fully joint optimization (including correspondences t_{np}).

5. Experiments

For the NRSfM unknowns given by θ (§ 3.2), we compare our variable correspondence based optimization: $\min_{\{\theta, t_{nuv}\}} E_{\text{gen}}$ against NRSfM with fixed correspondences: $\min_\theta E_{\text{gen}}$ (see (2)). We examine two classes: 'lilies' and 'clownfish'. Curve-based image evidence is employed to optimize similar objectives under different surface representations, class-based priors and varying amounts of occlusion. Class-specific curves are identified by first running a sub-pixel edge detector and edgel linker, and then manually selecting the corresponding edges. This is a relatively quick process, requiring a few clicks per image, but for a situation where thousands of images were to be labelled, partial automation would be desirable. While the optimization (and our implementation) handles any projection model, in the following experiments we use a 7-parameter similarity transform (un-normalized quaternion and translation).

5.1. Lilies

In the first class: 'lily', the open petal surface is assumed to be completely determined from the three "ribs" on the petal surface. We call this the "Wireframe Class Model" (WCM). Therefore the surface is reduced to a set of ribs—its underlying 'wireframe' representation—*i.e.* $u \in [1, U]$ but now $v \in \{1, 2, V\}, V = 3$ (for 3 petal ribs). Given the defining ribs, the rest of the surface is interpolated as shown in figs. 1,5,6. The notion of surface smoothness (§3.4) reduces to rib-based 3D curve smoothness. Computing second-order rib smoothness is equivalent to using *only* matrix operator $\mathtt{C}_{uu}$ on the individual ribs $X_n(u, v)$. The horizontal ($\mathtt{C}_{vv}$) and diagonal ($\mathtt{C}_{\text{uv}}$) smoothness terms are irrelevant across the ribs and are turned off (the ribs are not spatially next to each other on the surface, therefore are not expected to be smooth horizontally and diagonally on the parameter space). In addition to smoothness, vertices on the individual ribs must be as uniformly spaced as possible while adhering to image-based information. The tension term $\mathtt{C}_u$ is useful in this context. The final

$$\mathbf{e}^n_{\text{bending}} = \lambda C_{uu} \vec{\mathbf{X}}_n \tag{17}$$

$$\mathbf{e}^n_{\text{tension}} = \phi C_u \vec{\mathbf{X}}_n \tag{18}$$

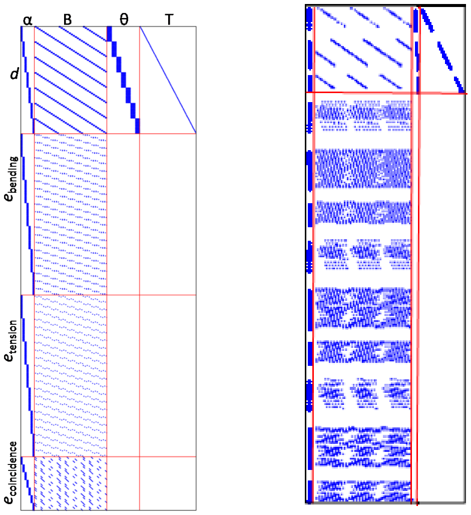
Smoothness in the above form (especially tension) can cause the ribs to shrink to singular points. Stated as in (2), a global optimum of the objective can be found by setting $\mathtt{B}_k = 0$ for all k, and choosing $\mathtt{P}_n$ to project the resulting point onto any point on the image curve, *e.g.* $\mathbf{w}_n(0, 0)$. The first problem is reduced if point correspondences are available. In our example, the tip and base of the petal are identifiable in many views, and can be included as conventional point constraints. However, image-based observation noise and annotation noise makes such constraints undesirable. A weaker, but nevertheless useful constraint is to encourage certain points on the curves to be coincident in 3D. This allows for the 3D rib tips to continue to be flexible and optimizable, promoting petal-like appearance, removing occurrence of singularity without explicit constraints. For example, given the point labelling in figs. 2,3, we add the following terms to the optimization:

$$\begin{aligned} E_{\text{pt}} = \sum_n \chi_{\text{top}} \left(\|\mathbf{X}_{n11} - \mathbf{X}_{n12}\|^2 + \|\mathbf{X}_{n11} - \mathbf{X}_{n13}\|^2 \right) \\ + \chi_{\text{bot}} \left(\|\mathbf{X}_{nU1} - \mathbf{X}_{nU2}\|^2 + \|\mathbf{X}_{nU1} - \mathbf{X}_{nU3}\|^2 \right) \end{aligned} \tag{19}$$

Combining these terms gives our primary objective:

$$E^{\text{wcm}}(\Theta, t_{11}, ..., t_{np}) = D_{\text{RP}} + E_{\text{pt}} + E_{\text{bending}} + E_{\text{tension}} \tag{20}$$

Observations: A collection of $N = 56$ "lily" (petal) photos were downloaded from Flickr, and manually annotated as described above, to produce a three-rib curve for each view. Optimizing (8) with fixed $t_n(u, v)$ and



(a) Lily (b) Clownfish

J (for a)	$\alpha(NK)$	$B(3KP)$	$\theta(7N)$	$T(NP)$
$\mathbf{d}$	$2PNK$	$6PNK$	$14NP$	$2NP$
$\mathbf{e}_{\text{bending}}$	$3PNK$	$\approx 9PNK$	0	0
$\mathbf{e}_{\text{tension}}$	$3PNK$	$\approx 9PNK$	0	0
$\mathbf{e}_{\text{pt}}$	$12PNK$	$72NK$	0	0

Figure 4. **Jacobian structure:** (a) 'Lily': the variables (for $(N = 4, K = 3, UV = 12)$) are plotted horizontally while the terms from the residual of E^{wcm} are vertical. The densities of the blocks for (a) are shown in the matrix below. (b) Clownfish: Jacobian for $(N = 4, K = 3, UV = 100)$ is on the right. Now, the vertical blocks to $D, \mathbf{e}_{\text{bending}}$ (Blocks for $e_{\text{pt}}, e_{\text{tension}}$ removed, but e_{bending} is larger due to surface-based smoothness). Additionally, a number of rows in (b) may be empty due to occluded vertices.

$U = 20$ unit-speed samples per rib ($\times V = 3$ ribs) is equivalent to a regularized version of NRSfM, and produces rather flat reconstructions (see fig. 5). For $N = 56$ images, $UV = 60$ and $K = 4$, the number of parameters $= NK + UVK3 + 7N = 1336$. Allowing for variable correspondences adds $NUV = 3360$ redundant variables, totalling to 4696. The Jacobian and its sparsity pattern can be seen in fig. 4 (a). The hyper-parameters $\{\lambda, \phi, \chi\}$ in this optimization are set empirically by visual reconstruction quality and requires little tuning in our experience; our WCM produces realistic 3D models regardless of the exact value of these hyper-parameters (unless scaled exorbitantly). Results are summed up in table 1 and figs. 3,5.

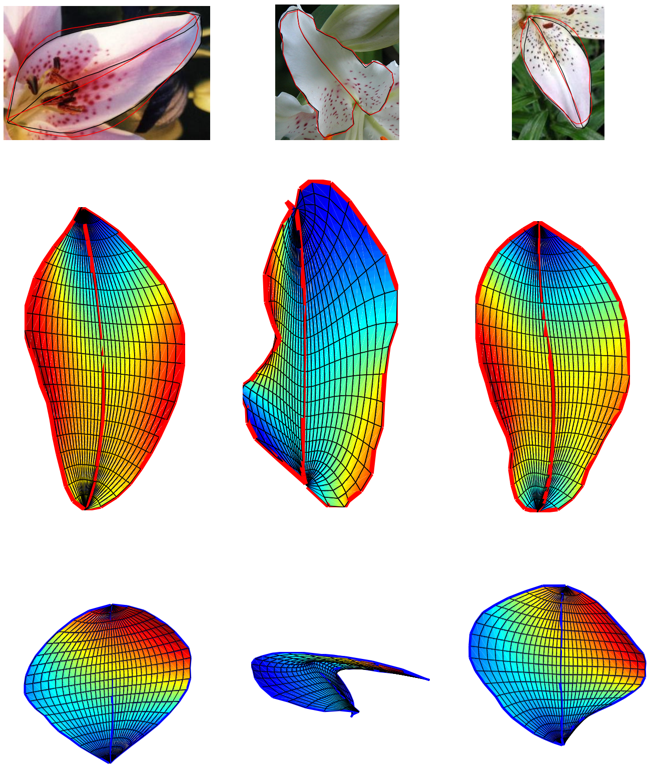


Figure 5. **WCM:** Row 1: Three lily images and annotations. Row 2: 3D wireframes (and interpolated petal surfaces) estimated by our WCM method ($K = 4$) using flexible correspondences are much more realistic. Our ribs are projected to Row 1 in red. Row 3: Our best results for standard NRSfM techniques, which produces mostly flat petals. Note: surface colour ranges from blue to red denoting vertex depth for visualization.

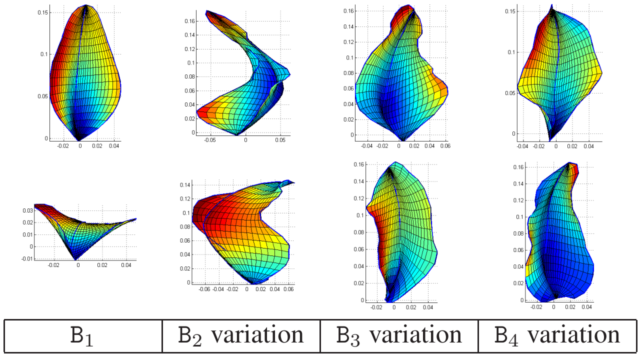


Figure 6. **Lily:** Modes of deformation. The bases are wireframe; surfaces are interpolated for ease of understanding.

5.2. Clownfish

We now consider a new case: "clownfish", adopting a slightly different approach from the WCM of §5.1. Instead, we solve for the entire $U \times V$ mesh as proposed in §3.1. Only a part of the clownfish surface corresponding to the observed image curves, is seen in each image. In addition to curve annotation, the user initializes approximate mapping of each curve to its parametric locations during the annotation (see fig. 3). This mapping only needs to be approximately correct for each curve, but the collective, relative ordering of all mappings on the surface must be reasonably correct. For the moment, we ignore the complex side fin to keep the topology simple at the current resolution. Since the observed data is limited to image-based curves, object parts corresponding to texture-less and non-silhouette regions are invisible to the method and considered "occluded" in that image. The individual vertex visibilities (see fig. 3) are important to the optimization. Reprojection error continues as (8). The clownfish is represented as a closed surface mesh ($U = 10, V = 10$) of cylindrical topology (see fig. 3). The regularization matrices are designed to incorporate the bending energy (9) and the topology [11]. The lily-specific E_{pt} (19) is removed. To deal with the whole surface, all terms of (9) are used as opposed to (17). Then the full objective is:

$$E^{\text{full}}(\Theta, t_{11}, ..., t_{np}) = D_{\text{RP}} + E_{\text{bending}} \tag{21}$$

Observations: A dataset of $N = 20$ images is collected off the web and annotated as described above (also see § 5). The number of vertices represented by each ω_{ni} is determined by their 3D parametric mapping. Optimizing (21) with fixed, unit-speed t_{nuv} yields the NRSfM solution, while allowing the correspondences to vary in (21) extends the method to find variable correspondences. For, $N = 20$ images, $UV = 100$, and $K = 4$ the number of parameters $= NK + UVK3 + 7N = 1420$. Interestingly, only 169 points of $NUV = 2000$ are ever visible; only 2 are seen in every image (see fig. 3 (e)). Allowing for variable correspondences adds $NUV = 2000$ additional redundant variables, totalling to 3420 (see Jacobian in fig. 4 (b)). Using bundle adjustment, the bases are built incrementally and separately for fixed and variable correspondences. This results in separate solutions with comparable function values as seen in table 1. At low λ ($1e-1$, weight on smoothness) and low bases ($K = 1$ here), variable correspondences produce improved 3D models and reduce E_{RP} and E_{full}. At higher λ ($1e4$) smoothness overtakes reprojection error. Numerically, NRSfM and our method produce similar results (see table 1), though visually our results continue to look better (see fig. 7).

K	NRSfM		Ours	
	E_{RP}	RMS E	E_{RP}	RMS E
$E = E^{\text{wcm}}$	Lily: $N = 56, P = 60, \lambda = 1e-2$			
1	12.16	0.4247	10.59	3.6974
2	9.22	0.3219	8.99	0.31
3	7.56	0.26	7.19	0.25
4	5.57	0.19	5.38	0.18
$E = E^{\text{full}}$	Clownfish: $\lambda = 1e4$			
1	4.9307	1.5671	4.9307	1.5671
2	3.4412	1.1527	3.8123	1.2632
3	2.2969	0.8271	2.7730	0.9834
	Clownfish: $\lambda = 1e-1$			
1	5.0982	1.6469	4.0107	1.3415
2	2.8267	0.8540	3.0968	1.1058
3	1.7469	0.5278	2.3942	0.9203

Table 1. Results for the competing methods: (i) **Lily**: The WCM method achieves a better minimum. Additionally, it optimizes reprojection error (reported per point) better too. (ii) **Clownfish:** Data being limited, our algorithm performs better (in terms of function value) at low regularization and low number of bases ($K = 1$). Despite higher function values at higher K, the visual reconstructions obtained by our method are more realistic (see fig. 7).

6. Summary

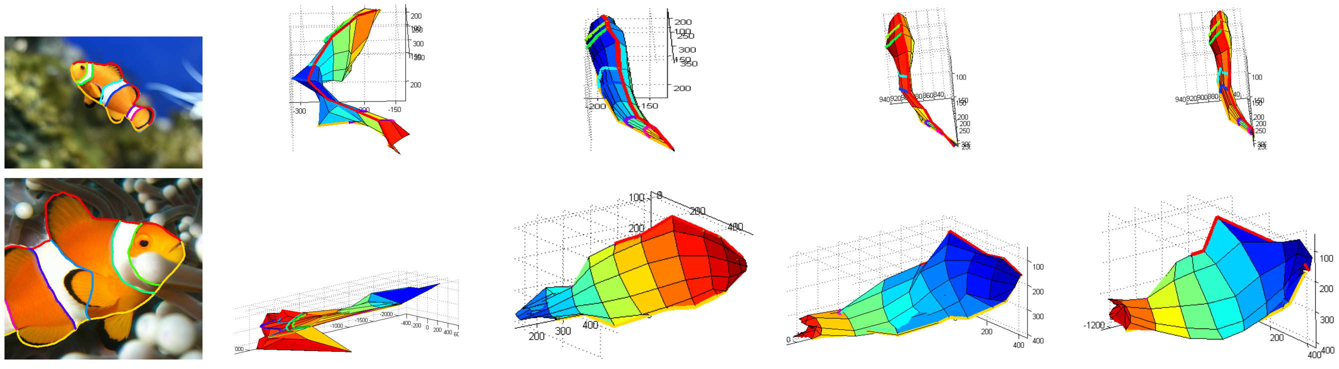
We have shown how a single bundle adjustment framework, built around curve features, allows a variety of 3D reconstruction from collections of similar, but distinct class instances despite the lack of point correspondences or temporal smoothness.

We first apply our method to find lily petal structure approximated by a rib-based wireframe. Image-curves representing rib projections are used to jointly estimate correspondences (up to a local minimum) along with the standard NRSfM variables. All vertices of the object are observable in a reasonably large dataset and we show significant improvement over existing techniques.

We then extend the method for "clownfish"—a topologically-cylindrical class—from partial image curve-based cues. While allowing for occlusion of unseen vertices in each image, correspondences are still jointly learnt with the rest of the variables. This is particularly interesting because in each image, most vertices happen to be invisible (fig. 3); those observed are often the same vertices. The results provide a captivating teaser for how far such methods can be pushed in the face of extreme occlusion and limited data.

When solved separately and incrementally (as described in § 4), the two competing methods–the fixed correspondence NRSfM and our variable correspondence based approach (using E^{wcm}, E^{full})–can land in different local optima of the complex objective function. Our optimization generally leads to more plausible optima than NRSfM, when both are initialized identically (barring inflection points). We lack ground truth to train and test these algorithms. Therefore, observation and annotation noise, inherent ambiguity in solutions, initialization issues and error in model assumptions affect the exact function value at the local optimum. Therefore, in addition to objective value, visual plausibility is an important benchmark in evaluating the final reconstructions.

In this paper, we have approached 3D deformable class reconstruction from a fresh perspective. We have not provided a closed-form, or factorization-based, algorithm, but rather used a carefully controlled bundle adjustment to prove the concept of 3D object class reconstruction. Note, however, that existing systems for structure and motion recovery, as well as recent NRSfM algorithms [16, 5, 12], all



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

Figure 7. Comparison: An image (a) and its reconstructions for $\lambda = 1e - 1$ (less smooth) with fixed (b) and variable (c) correspondences are shown. Also for $\lambda = 1e4$ (smoother) the fixed (d, with self-intersections) and variable (e) correspondence results are shown. Variable correspondences consistently result in more plausible reconstructions.

eschew factorization in favour of nonlinear minimization (whether expectation maximization or second-order methods), or of more realistic statistical or projection models. We hope that in the future, clever initializations will be found for these methods, but at this stage, we consider it valuable to have posed and examined the problem using the powerful tools available today.

An important extension to this method is the use of other obvious image cues *e.g.* silhouettes, and surface texture. We also hope to procure ground truth 3D exemplars for more comprehensive comparisons in the future.

Acknowledgements: We thank Thomas Minka for valuable help in discussions, implementation and proofreading. This work is supported by Microsoft Research, the Felix scholarship, ERC grant VisRec 228180 and 3D-COFORM.

References

[1] A. Bartoli, E. von Tunzelmann, and A. Zisserman. Augmenting images of non-rigid scenes using point and curve correspondences. In *Proc. CVPR*, Jun 2004. 2

[2] R. Berthilsson, K. Åström, and A. Heyden. Reconstruction of curves in $\mathbb{R}^3$, using factorization and bundle adjustment. In *Proc. ICCV*, volume 1, pages 674–679, 1999. 2

[3] M. Brand. Morphable 3D models from video. In *Proc. CVPR*, volume 2, pages 456–463, 2001. 1

[4] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *Proc. CVPR*, volume 2, pages 690–696, 2000. 1

[5] A. Del Bue. A factorization approach to structure from motion with shape priors. In *Proc. CVPR*, 2008. 1, 2, 7

[6] A. W. Fitzgibbon. Robust registration of 2D and 3D point sets. In *Proc. BMVC.*, pages 662–670, 2001. 4

[7] W. E. L. Grimson. *From Images to Surfaces: A Computational Study of the Human Early Visual System*. MIT Press, 1981. 4

[8] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 2

[9] J. Y. Kaminski and A. Shashua. Multiple view geometry of general algebraic curves. *IJCV*, 56(3):195–219, 2004. 2

[10] H. Martinsson, F. Gaspard, A. Bartoli, and J.-M. Lavest. Energy-based reconstruction of 3D curves for quality control. In *Proc. EMMCVPR*, 2007. 2

[11] M. Prasad, A. Zisserman, and A. W. Fitzgibbon. Single view reconstruction of curved surfaces. In *Proc. CVPR*, volume 2, pages 1345–1354, June 2006. 4, 7

[12] V. Rabaud and S. Belongie. Re-thinking structure from motion. In *Proc. CVPR*, 2008. 2, 7

[13] C. Schmid and A. Zisserman. The geometry and matching of curves in multiple views. In *Proc. ECCV*, pages 394–409. Springer-Verlag, Jun 1998. 2

[14] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. *ACM Trans. Graph.*, 25(3):835–846, 2006. 1

[15] R. Szeliski. Fast surface interpolation using hierarchical basis functions. *IEEE PAMI*, 12(6):513–528, 1990. 4

[16] L. Torresani, A. Hertzmann, and C. Bregler. Non-rigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE PAMI*, 30(5):878–892, 2008. 1, 2, 7

[17] W. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment: A modern synthesis. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, LNCS, pages 298–375. Springer Verlag, 2000. 2

[18] L. Zhang, G. Dugas-Phocion, J. Samson, and S. Seitz. Single view modeling of free-form scenes. In *Proc. CVPR*, pages I:990–997, 2001. 4