Combining Discriminative and Generative Methods for 3D Deformable Surface and Articulated Pose Reconstruction

Mathieu Salzmann
EECS & ICSI, UC Berkeley
salzmann@icsi.berkeley.edu

Raquel Urtasun
TTI, Chicago
rurtasun@ttic.edu

Abstract

Historically non-rigid shape recovery and articulated pose estimation have evolved as separate fields. Recent methods for non-rigid shape recovery have focused on improving the algorithmic formulation, but have only considered the case of reconstruction from point-to-point correspondences. In contrast, many techniques for pose estimation have followed a discriminative approach, which allows for the use of more general image cues. However, these techniques typically require large training sets and suffer from the fact that standard discriminative methods do not enforce constraints between output dimensions. In this paper, we combine ideas from both domains and propose a unified framework for articulated pose estimation and 3D surface reconstruction. We address some of the issues of discriminative methods by explicitly constraining their prediction. Furthermore, our formulation allows for the combination of generative and discriminative methods into a single, common framework.

1. Introduction

Recent advances in monocular non-rigid surface reconstruction have focused on designing formulations of the problem that are easier to optimize (e.g., convex) [22, 10, 14]. Typically, the shape is parameterized in terms of the 3D vertex coordinates of a mesh, and the problem is addressed in a generative framework where an image likelihood is minimized under constraints on the distances between neighboring vertices of the mesh. This parameterization has the advantage of facilitating convex formulations of the likelihood as well as relaxations of the constraints [14]. However, these techniques rely on dense point-to-point correspondences, which might be difficult to obtain, particularly when the surface is poorly textured, and cannot make use of non-registered features.

In articulated pose estimation, many techniques have focused on learning a mapping from image observations to 3D poses [8, 13, 1, 17, 19]. The main strength of these discriminative approaches is that they do not require point-to-point correspondences, and can take advantage of any type of image representation as long as a kernel between pairs of images can be computed. However, since typical discriminative methods assume independence of the output dimensions given the inputs, important constraints, such as limb lengths, are often violated when modeling the pose in terms of the joint coordinates of a skeleton. As illustrated in Fig. 1(a) for the case of a non-rigid surface, this may yield unrealistic poses. To overcome this weakness the human body is typically parameterized in terms of joint angles [1, 19]. However, designing such a parameterization for deformable surfaces is not straightforward.

In any event, an inherent problem of discriminative methods is that they require large training sets to account for the high dimensionality and variability of the pose space. This might be one of the reasons why they have not yet been used for non-rigid 3D surface reconstruction, and have mainly been applied to a restrictive set of activities for human pose estimation. As a consequence, while discriminative approaches have proved very effective for classification tasks such as object recognition, they are often outperformed by generative ones for regression tasks such as pose estimation [2, 20]. Unfortunately, the success of generative techniques heavily depends on having a good initialization, since they typically rely on non-convex objective functions. Discriminative methods could be employed to obtain this initialization. However, due to the lack of training data, or the typical output independence assumption, they are often not precise enough, and the non-convex generative methods get trapped in local minima. This suggests that a more principled combination of generative and discriminative methods would be key for the success of pose estimation and non-rigid surface reconstruction. While this was also observed by [18, 9, 16], the first approach relies on the generative method only for learning, and the other two use the discriminative technique only for initialization.

In this paper we combine the findings of the human pose estimation and non-rigid shape recovery domains and show that both problems can be solved within the same framework. Our approach addresses some of the issues of discriminative methods by introducing explicit constraints and forcing the prediction to satisfy them. In particular, we consider the case of distance constraints between neighboring 3D points on a mesh or on a human skeleton (i.e., joints).

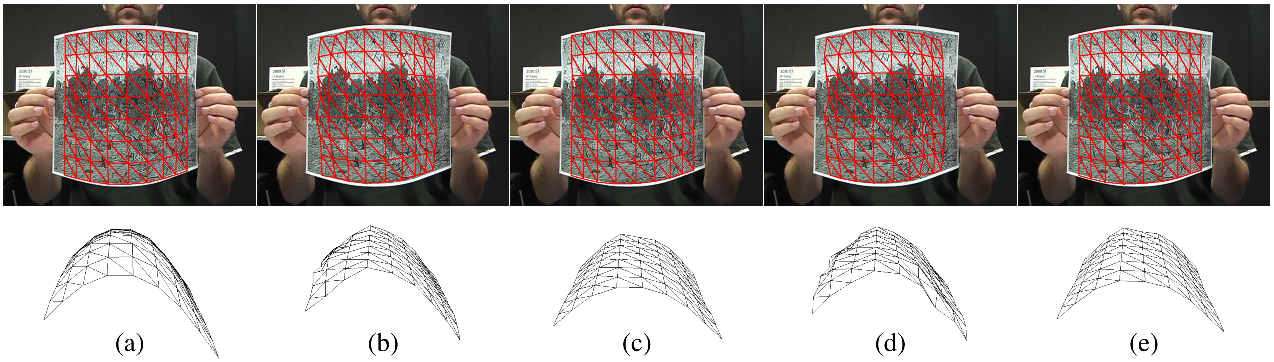


Figure 1. **Reconstructing a piece of paper from monocular images.** Top row: Recovered mesh reprojected on the input image. Bottom row: Side view of the same mesh. Results were obtained with (a) the original predictor, (b) the constrained predictor, (c) the constrained predictor with an image likelihood, (d,e) same as (b,c) but when optimizing $\mathbf{k}_*$. Note that the predictor's result reprojects correctly but has noticeably stretched, whereas using the constraints only does not ensure a correct reprojection.

This lets us combine discriminative and generative methods into a common formulation that, for image-based squared loss functions, simply involves iteratively solving a set of linear equations.

The contributions of this paper can be summarized as follows: We propose a novel approach to incorporating explicit constraints in discriminative methods. We combine discriminative and generative methods within a single framework. We present a unified formulation of articulated pose estimation and non-rigid shape reconstruction and demonstrate the effectiveness of our approach on synthetic and real monocular images.

2. Constrained Discriminative Regression

In this section we introduce our approach to incorporating explicit relationships between the output dimensions in discriminative methods. In particular, we consider the case of estimating the 3D pose of a human represented as a skeleton, and the case of reconstructing the 3D shape of a non-rigid surface modeled as a triangulated mesh. We introduce equality constraints that represent fix distances between pairs of joints or between mesh vertices. In the remainder of this paper, we will use the word *pose* to refer to a human pose as well as a surface shape.

Let $\mathbf{x} \in \Re^Q$ be a random variable representing an input observation (e.g., image features), and $\mathbf{y} \in \Re^D$ the associated output (e.g., pose). Discriminative methods do not directly model the joint statistics of these two random variables and focus on estimating the possibly non-linear mapping $\mathbf{f}$ between them, such that

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) + \epsilon \; , \tag{1}$$

where $\mathbf{f} = [f_1, \cdots, f_D]$, and each f_i predicts a single output dimension i, assumed to be corrupted by i.i.d. noise ϵ_i.

Given a set of i.i.d. pairs, $(\mathbf{x}^1, \mathbf{y}^1), \cdots, (\mathbf{x}^N, \mathbf{y}^N)$, sampled from the joint distribution $p(\mathbf{x}, \mathbf{y})$, an estimate of $\mathbf{f}$, $\hat{\mathbf{f}}$, can be learned by empirical risk minimization. Different loss functions have been proposed (e.g., squared loss, hinge loss), each of which yields a different discriminative technique (e.g., least-squares regression, SVMs).

Typically, when $\mathbf{y}$ is multi-dimensional (i.e., $D > 1$), its dimensions are assumed to be independent given the inputs [1, 19]; A different regressor is trained for each of the output dimensions, thus ignoring their dependencies. As a consequence, the prediction $\hat{\mathbf{f}}(\mathbf{x}_*)$ for a new input $\mathbf{x}_*$ might not satisfy these constraints. This problem is illustrated in Fig. 1(a), where distance constraints between the vertices of a mesh are violated, thus yielding a stretched surface. We now show how to incorporate explicit constraints into discriminative methods to prevent this from happening.

2.1. Distance Preservation Constraints

Let $\mathbf{y} \in \Re^{3N_p}$ be the vector of 3D coordinates of the N_p points that define a pose. For human pose recovery, natural constraints arise from the fact that the skeleton is a kinematic tree, and the distance between a parent node and its children should remain constant as the person moves. In the case of inextensible surfaces, it has been shown that constraining the length of the mesh edges to remain constant effectively disambiguates the reconstruction process [5, 14].

Let $\mathcal{E}$ be the set of N_e links between 3D points whose length should remain constant. Given a new input observation $\mathbf{x}_*$ and the prediction of a discriminative method $\hat{\mathbf{f}}(\mathbf{x}_*)$, constrained pose estimation can be formulated as solving the optimization problem

$$\begin{aligned} \underset{\mathbf{y}}{\text{minimize}} \quad & ||\hat{\mathbf{f}}(\mathbf{x}_*) - \mathbf{y}||_2^2 \\ \text{subject to} \quad & ||\mathbf{y}_k - \mathbf{y}_j||_2^2 = l_{j,k}^2 \; , \; \forall (j,k) \in \mathcal{E} \; , \end{aligned} \tag{2}$$

where $\mathbf{y}_i$ is the subvector of $\mathbf{y}$ containing the i-th 3D point, and $l_{j,k}^2$ is the fixed squared distance between points j and k. Because of the distance equality constraints, this formulation is non-convex. However, Shen et al. [14] showed that for 3D shape recovery these distance constraints could be incorporated in a generative approach and enforced by iteratively solving a linear system. Here, we follow a similar

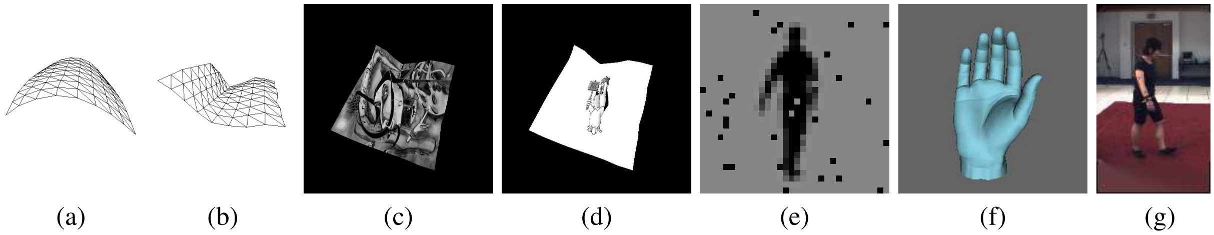


Figure 2. **Samples from our datasets.** (a) Mesh corresponding to a deformed piece of cardboard reconstructed with a motion capture system. (b) Synthetically generated inextensible mesh. (c) Image generated by texturing the mesh in (b). (d) Similar image obtained with a more uniform texture. (e) Noisy silhouette of walking person [1]. (f) Input image for a synthetic hand dataset [19]. (g) Image from the HumanEva dataset [15] registered with the method of [7].

idea in the context of discriminative methods and show that it still involves iteratively solving a linear system, but with different equations.

Let $\mathbf{y}_t$ be the estimate of the 3D shape at iteration t, which does not satisfy the distance constraints. Obtaining a shape that satisfies our constraints can be formulated as finding a small displacement $\delta\mathbf{y}_t$ such that

$$||\mathbf{E}_{j,k}(\mathbf{y}_t + \delta\mathbf{y}_t)||_2^2 = l_{j,k}^2 \ , \ \forall (j,k) \in \mathcal{E} \ , \qquad (3)$$

where $\mathbf{E}_{j,k}$ is the $3 \times 3N_p$ matrix encoding the distance constraint for the edge linking point j to point k. Expanding the previous equation yields

$$\mathbf{y}_t^T\mathbf{E}_{j,k}^T\mathbf{E}_{j,k}\mathbf{y}_t+2\mathbf{y}_t^T\mathbf{E}_{j,k}^T\mathbf{E}_{j,k}\delta\mathbf{y}_t+\delta\mathbf{y}_t^T\mathbf{E}_{j,k}^T\mathbf{E}_{j,k}\delta\mathbf{y}_t = l_{j,k}^2 \ . \qquad (4)$$

Similarly to [14], if we assume that the current estimate $\mathbf{y}_t$ is close to the true solution, and therefore that $\delta\mathbf{y}_t$ is small, we can neglect the quadratic term in $\delta\mathbf{y}_t$. Doing so and grouping all distance constraints yields the linear system

$$\mathbf{F}_t\delta\mathbf{y}_t = \mathbf{g}_t \ , \qquad (5)$$

where the i-th rows of $\mathbf{F}_t$ and $\mathbf{g}_t$ can be computed as

$$\mathbf{F}_{t,i} = 2\mathbf{y}_t^T\mathbf{E}_{j,k}^T\mathbf{E}_{j,k}, \quad \mathbf{g}_{t,i} = l_{j,k}^2 - \mathbf{y}_t^T\mathbf{E}_{j,k}^T\mathbf{E}_{j,k}\mathbf{y}_t \ .$$

Note that this formulation is equivalent to performing a first order Taylor expansion and therefore generalizes to any equality constraints. Since triangulated meshes and human skeletons have more 3D coordinates than edges, the system of Eq. 5 has more unknowns than equations. Therefore, it yields the family of solutions

$$\delta\mathbf{y}_t = \mathbf{F}_t^+\mathbf{g}_t + \mathbf{V}_t^T\boldsymbol{\gamma}_t \ , \qquad (6)$$

where $\mathbf{F}_t^+$ is the pseudo-inverse of $\mathbf{F}_t$, $\mathbf{V}_t$ is the matrix containing the last $(3N_p - N_e)$ right singular vectors of $\mathbf{F}_t$ which have zero-valued singular values, and $\boldsymbol{\gamma}_t$ is the $(3N_p - N_e)$ dimensional vector of remaining unknowns.

Given these new unknowns that implicitly minimize the violation of the constraints, we can re-write Eq. 2 as

$$\underset{\boldsymbol{\gamma}_t}{\text{minimize}} \, ||\hat{f}(\mathbf{x}_*) - (\mathbf{y}_t + \mathbf{F}_t^+\mathbf{g}_t + \mathbf{V}_t^T\boldsymbol{\gamma}_t)||_2^2 \ . \qquad (7)$$

This is now a convex optimization problem, whose minimum can be obtained in closed-form by solving a linear system in the least-squares sense.

Since solving the problem of Eq. 7 only once may yield a $\delta\mathbf{y}_t$ too large to make the quadratic term of Eq. 4 negligible, we iterate until the maximum constraint violation is less than a pre-defined threshold or a maximum number of iterations T has been reached. At each iteration t, we compute $\mathbf{F}_t$, $\mathbf{g}_t$ and $\mathbf{V}_t$, solve the problem of Eq. 7, and update $\mathbf{y}_t$ with the resulting $\delta\mathbf{y}_t$. In practice, we initialize $\mathbf{y}_0$ with the prediction of the discriminative method.

2.2. Stronger Dependencies on the Predictor

While the approach described above lets us constrain the outputs of a discriminative method, it depends on the predictor only through its fixed prediction $\hat{f}(\mathbf{x}_*)$. To make better use of the predictor, we rely on the *Representer theorem* [12] which states that, if $\hat{f}$ is the minimizer of an L2-regularized empirical loss function $L : \Re^Q \rightarrow \Re$, then

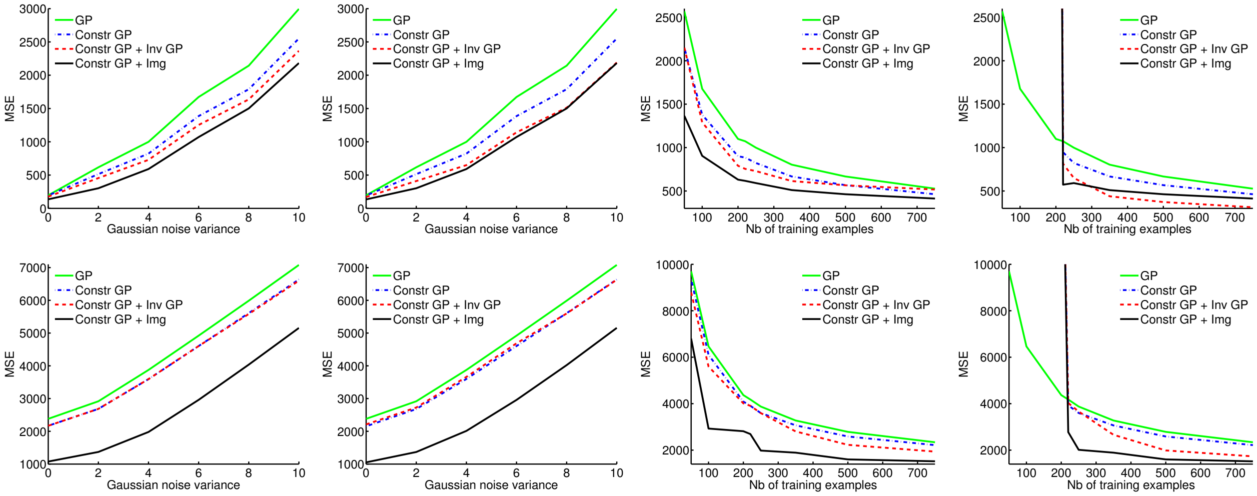
$$\hat{f}(\mathbf{x}_*) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}_*) = \alpha\mathbf{k}_* \qquad (8)$$

with k a kernel function, and $\alpha = [\alpha_1, \cdots, \alpha_N]$. For multi-dimensional outputs, assuming that the output dimensions are independent, this yields one α per $\hat{f}_j \ , \ 1 \leq j \leq D$, which can be grouped to form a matrix $\boldsymbol{\alpha}$. Therefore, to rely more heavily on the predictor, we propose to exploit the basis $\boldsymbol{\alpha}$ defined by the representer theorem and learned at training. To this end, instead of optimizing with respect to the shape $\mathbf{y}$, we search for the vector $\mathbf{k}_*$ that defines the predicted pose by minimizing $||\hat{\mathbf{f}}(\mathbf{x}_*) - \boldsymbol{\alpha}\mathbf{k}_*||_2^2$ subject to the distance constraints.

Since the prediction in Eq. 8 is a linear function of $\mathbf{k}_*$, we can use a similar iterative process as before, and search for a small variation $\delta\mathbf{k}_{*,t}$ around a current estimate $\mathbf{k}_{*,t}$ such that the new estimate satisfies the linearized distance constraints. At each iteration, the optimal $\delta\mathbf{k}_{*,t}$ can be obtained by solving the linear system $\mathbf{F}_t\delta\mathbf{k}_{*,t} = \mathbf{g}_t$, where the rows of $\mathbf{F}_t$ and $\mathbf{g}_t$ are now defined as

$$\mathbf{F}_{t,i} = 2\mathbf{k}_{*,t}^T\boldsymbol{\alpha}^T\mathbf{E}_{j,k}^T\mathbf{E}_{j,k}, \ \ \mathbf{g}_{t,i} = l_{j,k}^2 - \mathbf{k}_{*,t}^T\boldsymbol{\alpha}^T\mathbf{E}_{j,k}^T\mathbf{E}_{j,k}\boldsymbol{\alpha}\mathbf{k}_{*,t}.$$

As the dimensionality of $\mathbf{k}_{*,t}$ depends on the number of training examples N, we have multiple solutions only when $N > N_e$. In that case, we can describe the family of solutions in terms of new unknowns $\boldsymbol{\gamma}_t$ similarly as in Eq. 6,



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

Figure 3. **Reconstructing a non-rigid surface with 2D locations as input.** Top: Deforming piece of cardboard of Fig. 2(a). Bottom: Inextensible meshes of Fig. 2(b). Average MSE as a function of (a,b) noise variance, and (c,d) number of training examples. (a,c) were obtained by optimizing $\mathbf{y}$ and (b,d) by optimizing $\mathbf{k}_*$.

and re-write the problem of Eq. 2 as

$$\underset{\boldsymbol{\gamma}_t}{\text{minimize}}\ ||\hat{\mathbf{f}}(\mathbf{x}_*) - \boldsymbol{\alpha} \cdot (\mathbf{k}_{*,t} + \mathbf{F}_t^{+}\mathbf{g}_t + \mathbf{V}_t^T \boldsymbol{\gamma}_t)||_2^2 \ . \quad (9)$$

This problem can again be solved in closed-form. We iterate over t until all constraints are satisfied or a maximum number of iterations is reached.

3. Combining Discriminative and Generative

One drawback of the methodology described in the previous section is that it only uses image information through the prediction of the discriminative method. Therefore, the recovered pose will satisfy the constraints, but may have drifted away from the pose depicted in the image, as illustrated by Fig. 1(b,d). In this section, we propose to make use of the image more explicitly, while enforcing the constraints and regularizing the solution to be close to the discriminative prediction. To this end, we rely on the formulation introduced in the previous section, and iteratively linearize our distance constraints to solve the problem

$$\underset{\boldsymbol{\gamma}_t}{\text{minimize}}\ \ \mathcal{L}(\cdot, \boldsymbol{\gamma}_t) + \lambda ||\hat{\mathbf{f}}(\mathbf{x}_*) - \mathbf{s}(\boldsymbol{\gamma}_t)||_2^2 \ , \quad (10)$$

where $\mathcal{L}(\cdot, \boldsymbol{\gamma}_t)$ is a loss function that depends on the image, and λ is a weight that sets the relative influence of both terms. As shown in the experiments, our method is not very sensitive to the value of λ. As in the previous section, we can solve the problem in terms of the pose $\mathbf{y}_t$ by setting $\mathbf{s}(\gamma_t) = \mathbf{y}_t + \mathbf{F}_t^{+}\mathbf{g}_t + \mathbf{V}_t^T \gamma_t$, or in terms of the kernel $\mathbf{k}_{*,t}$, by setting $\mathbf{s}(\gamma_t) = \boldsymbol{\alpha} \cdot (\mathbf{k}_{*,t} + \mathbf{F}_t^{+}\mathbf{g}_t + \mathbf{V}_t^T \gamma_t)$.

In particular, we propose two different approaches to combining the discriminative prediction with the image information: via learning an inverse mapping from pose to image features, and via a generative approach where an image likelihood is minimized. In the remainder of this section, we present these two approaches.

3.1. Using an Inverse Mapping

One possible way of relating the pose that we optimize to the original input is by making use of discriminative methods. Similarly as we learned a mapping from image observations to pose, we can learn an inverse mapping from pose to image features.

Given the learned inverse regressor $\hat{\mathbf{h}} : \mathbf{y} \rightarrow \mathbf{x}$, we can define the loss function in Eq. 10 as

$$\mathcal{L}(\mathbf{x}_*, \boldsymbol{\gamma}_t) = ||\hat{\mathbf{h}}(\mathbf{s}(\boldsymbol{\gamma}_t)) - \mathbf{x}_*||_2^2, \quad (11)$$

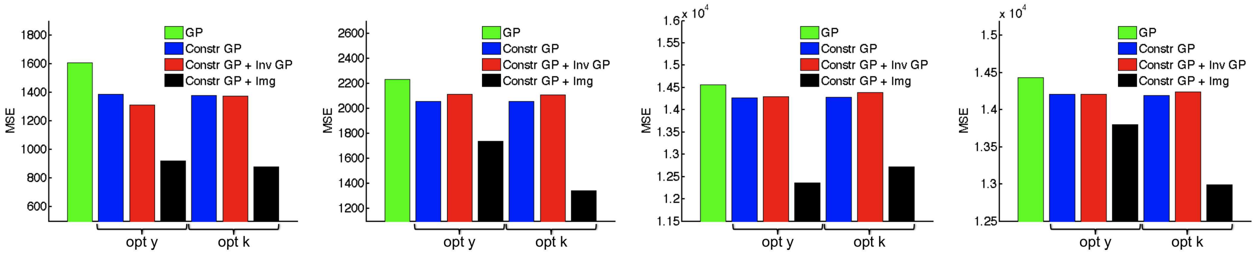
which describes the squared distance between the inverse prediction and the input image features. Note that, in general, this yields a non-convex optimization problem, which we solve using a quasi-newton solver. However, thanks to our constraints, our initial estimate is sufficiently good to only need a few iterations for accurate prediction. In practice, since linearizing the constraints involves an iterative scheme, we initialize $\boldsymbol{\gamma}_t$ at each step to the value that yields the constrained shape closest to the predicted pose.

3.2. Using an Image Likelihood

A more classical way of making use of the image observations is to employ an image likelihood. To this end, we encourage the optimized pose to correctly model the image information while remaining close to the discriminative prediction $\hat{\mathbf{f}}(\mathbf{x}_*)$. Here, we present the two image likelihoods used for our experiments. Note that other image likelihoods could also be utilized.

3.2.1 Minimizing the Reprojection Error

A standard approach to recovering the shape of a deformable surface is to minimize the reprojection error of points on the surface [22, 14, 10]. Similarly, for articulated



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

Figure 4. **Non-rigid reconstruction from Pyramid HOG.** Average MSE for (a,c) a well-textured piece of cardboard and inextensible mesh (e.g., Fig. 2(c)), and (b,c) poorly-textured surfaces (e.g., Fig. 2(d)). Legends below the plots indicate when we optimize $\mathbf{y}$ or $\mathbf{k}_*$.

pose estimation one can minimize the reprojection error between the 3D joints and the 2D joint locations estimated from the image [20, 4]. To this end, let $\mathbf{x}$ be the vector of 2D locations of the 3D points that define a pose. Given a new vector of 2D locations $\mathbf{x}_*$, we seek to recover the pose $\mathbf{s}(\boldsymbol{\gamma}_t)$ that reprojects near $\mathbf{x}_*$ and remains close to $\hat{\mathbf{f}}(\mathbf{x}_*)$.

Let us assume that the camera is calibrated, with $\mathbf{A}$ the known matrix of internal parameters. Without loss of generality, let us assume that $\mathbf{s}(\boldsymbol{\gamma}_t)$ is defined in the camera coordinate system. We can express the fact that a particular vertex $\mathbf{s}_i(\boldsymbol{\gamma}_t)$ should reproject at the 2D location $\mathbf{x}_{*,i}$ as

$$\mathbf{A}s_i(\gamma_t) = d_{i,t} \begin{bmatrix} \mathbf{x}_* \\ 1 \end{bmatrix} , \tag{12}$$

where $d_{i,t}$ is a scalar accounting for depth. This formulation yields two linear equations per point, and the equations for all 3D points can be grouped into a single linear system [10]. This lets us define the loss function of Eq. 10 as

$$\mathcal{L}(\mathbf{x}_*, \boldsymbol{\gamma}_t) = ||\mathbf{M}_t \cdot \mathbf{s}(\boldsymbol{\gamma}_t)||_2^2 , \tag{13}$$

where $\mathbf{M}_t$ is the $2N_p \times 3N_p$ matrix encoding the projection equations at iteration t. Unlike constraining the inverse mapping, this image likelihood is convex. Furthermore, it lets us re-formulate the problem of Eq. 10 as the least-squares solution to a linear system, which can be obtained in closed-form. This is due to the fact that, whether we use $\mathbf{y}_t$ or $\mathbf{k}_{*,t}$, $\mathbf{s}(\boldsymbol{\gamma}_t)$ is a linear function of $\boldsymbol{\gamma}_t$. Therefore, at each iteration t, we only need to solve a linear system.

3.2.2 A More General Likelihood Function

For non-rigid surfaces, when there is not enough texture to be able to recover the 2D locations of the vertices, we propose to rely on template matching and on surface boundary information. Note that, provided with a more accurate model than just a skeleton [16], this could also be applied to human pose estimation.

For template matching, given a reference image in which we know the shape of the surface and the camera projection matrix, the template $\mathbf{T}$ is obtained by sampling the barycentric coordinates of the mesh facets and collecting their corresponding intensities. The same process is used to obtain the intensities $\mathbf{J}(\mathbf{I}, \boldsymbol{\gamma}_t)$ in the input image $\mathbf{I}$ from the optimized shape. Matching is then done by maximizing the normalized cross-correlation between $\mathbf{T}$ and $\mathbf{J}$.

For boundaries, we first detect image edges using Canny's algorithm. We then project the current shape estimate into the edge image, sample its boundary, and, for each sample, look for corresponding image edge points along the normal direction to the mesh boundary. We then minimize the distance between the sampled boundary points and their corresponding image edge points. To be more robust, we allow for multiple candidates for each boundary point.

In this framework, the loss in Eq. 10 can be written as

$$\mathcal{L}(\mathbf{I}, \gamma_t) = -\Psi(\mathbf{T}, \mathbf{J}(\mathbf{I}, \gamma_t)) + \lambda_e \sum_{i=1}^{N_b} \sum_{j=1}^{N_c(i)} \|\mathbf{u}_{i,j} - \mathbf{e}_i(\gamma_t)\|^2 ,$$

where $\Psi(\cdot)$ is the normalized cross-correlation function, N_b is the number of sampled boundary points, $\mathbf{e}_i$ denotes the boundary point projected in the image, $N_c(i)$ is the number of edge candidates for point i, and $\mathbf{u}_{i,j}$ is the corresponding image measurement. λ_e is a weight that sets the relative influence of the two terms. As with the inverse regressor, this yields a non-convex objective function which we minimize using a quasi-newton solver. At each step of the iterative linearization scheme, we recompute the edge candidates and initialize $\boldsymbol{\gamma}_t$ to the value yielding the constrained pose closest to the discriminative prediction $\hat{\mathbf{f}}(\mathbf{x}_*)$.

4. Experimental Evaluation

In this section, we compare the performance of the different approaches proposed in this paper with the original discriminative method in the context of non-rigid shape reconstruction and articulated pose estimation.

For non-rigid surfaces, we used two types of output data: The reconstructed deformations of a piece of cardboard obtained with an infrared optical motion capture system, and inextensible meshes synthetically generated by randomly setting the angles between their facets [11]. Both types of meshes are 9×9 square grids of 16cm of side. As inputs, we used either the 2D vertex locations obtained with a known camera and corrupted with i.i.d additive Gaussian noise, or Pyramid HOG (PHOG) features computed from images obtained by texturing the meshes. Note that the 2D vertex locations could typically be obtained by methods such as [6, 21]. Fig. 2(a-d) depicts output and input samples from these datasets. Additionally, we applied our technique to real sequences of two different surfaces, and,

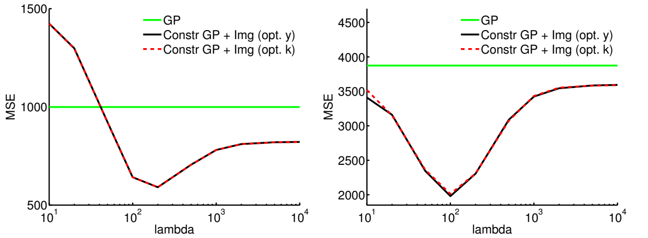


Figure 5. **Influence of λ in Eq. 10**. We plot in log-lin axes the MSE as a function of λ. For both the cardboard (left) and the inextensible meshes (right), our method outperforms the GP for a very large range of λ. Note that the curves for $\mathbf{y}$ and $\mathbf{k}_*$ are superposed and are best seen in color.

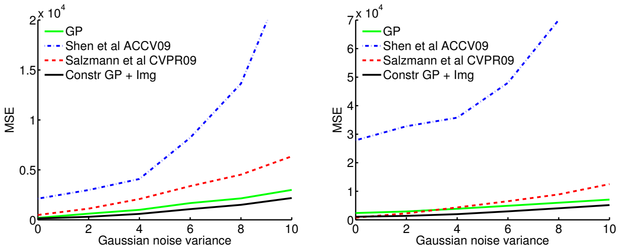


Figure 6. **Comparison against [14] and [10] for our two datasets.** MSE as a function of the noise variance. In both cases, our approach outperforms these techniques.

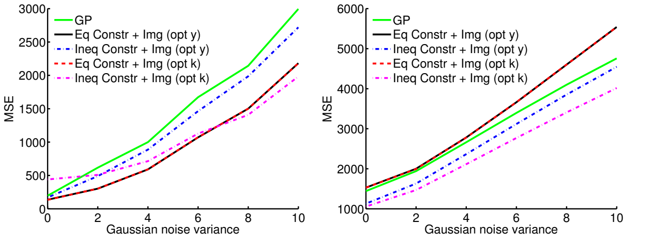


Figure 7. **Using distance inequalities [10] instead of equalities.** Left: For the deformations of a piece of cardboard both constraints give similar results. Right: For a more flexible surface, inequalities outperform distance equalities. As some curves are superposed, results are best seen in color.

as inputs, used the 2D vertex locations obtained by tracking the surface in 2D using template matching.

For articulated poses, we used three different datasets: A full body pose dataset [1] and a hand dataset [19] generated using Poser^{TM}, and HumanEva [15]. In all cases, the outputs of our discriminative method were taken as the 3D coordinates of 19 joints for the body and 17 for the hand. For the Poser body dataset, we used binary silhouettes as inputs, and simulated noise by switching the values of randomly selected pixels. For the Poser hand dataset, our inputs were taken as either spatial pyramids of steerable filters, of SIFT, or PHOG features computed from the images. For HumanEva, which consists of motion captured human body poses synchronized with videos, we used the registered walking images of [7] to compute spatial pyramids of SIFT and PHOG features, which we took as inputs. To test our method on a less restrictive set of activities, we also used the 2D joint locations for walking, jogging and boxing as inputs. These locations could typically be obtained in a similar manner as in [20, 4]. For both cases, we used the poses of a single subject. Samples are shown in Fig. 2(e-f).

The errors shown in our plots were computed as follows: We took the mean over the test examples of the sum of squared differences between our reconstructions and the ground-truth. We then averaged this quantity over 10 partitions of the data into training and test examples, and refer to the resulting value as mean squared error (MSE).

4.1. Our Choice of Predictor

In this paper we use Gaussian processes (GPs) [3] to learn both the direct and the inverse mappings. The use of GPs is particularly interesting since $\boldsymbol{\alpha}$ can be computed in closed form as $\boldsymbol{\alpha} = \mathbf{Y}^T\mathbf{K}^{-1}$, where $\mathbf{Y} \in \Re^{N \times D}$ is the matrix of training outputs (e.g., poses), and $\mathbf{K}$ is the covariance matrix whose entries are formed by evaluating the kernel function $k(\mathbf{x}^i, \mathbf{x}^j)$ on the training inputs. The prediction on which our method relies is taken as the mean prediction of the GP, $\hat{\mathbf{f}}(\mathbf{x}_*) = \boldsymbol{\alpha}\mathbf{k}_*$. For the inverse mapping, $\boldsymbol{\alpha}_x = \mathbf{X}^T\mathbf{K}_x^{-1}$, where $\mathbf{X} \in \Re^{N \times Q}$ is the matrix of training input observations (e.g., image features), and $\mathbf{K}_x$ is the covariance matrix evaluated on the training outputs, with entries $k_x(\mathbf{y}^i, \mathbf{y}^j)$. The prediction can be computed as $\hat{\mathbf{h}}(\mathbf{y}_*) = \boldsymbol{\alpha}_x\mathbf{k}_{x,*}$. For both mappings, our kernel was taken to be the sum of an RBF kernel and a bias.

4.2. Non-Rigid Surface Reconstruction

We first present the results of our approach on the deformable surfaces datasets. Fig. 3 depicts the results obtained by the original GP, the constrained GP, the constrained GP used in conjunction with the inverse mapping, and the constrained GP used in conjunction with an image likelihood. Plots (a) and (b) depict the MSE as a function of the variance of the noise for 250 training examples, and plots (c) and (d) as a function of the number of training examples for a fixed noise variance of 4, which corresponds to what can be expected from 2D non-rigid registration techniques. Fig. 3(a) and (c) correspond to optimizing $\mathbf{y}$, and (b) and (d) to optimizing $\mathbf{k}_*$. Optimizing $\mathbf{k}_*$ when there are less training examples than the number of constraints (208 in this case) yields an overconstrained problem, which explains the large errors. Fig. 4 shows the MSE obtained from PHOG features for the different types of data when the surface is well-textured (a,c) and when it is poorly-textured (b,d). Note that constraining the discriminative method and combining it with a generative approach significantly reduces the reconstruction error.

We then studied the influence of the weight λ of Eq. 10 on our reconstructions in the case where the inputs are the 2D vertex locations. Fig. 5 depicts the MSE as a function of λ in log-scale. Note that the curves are relatively flat, and that we outperform the GP for a wide range of λ.

We also compared our approach against two state-of-the-art techniques [14, 10]. The first one relies on the same

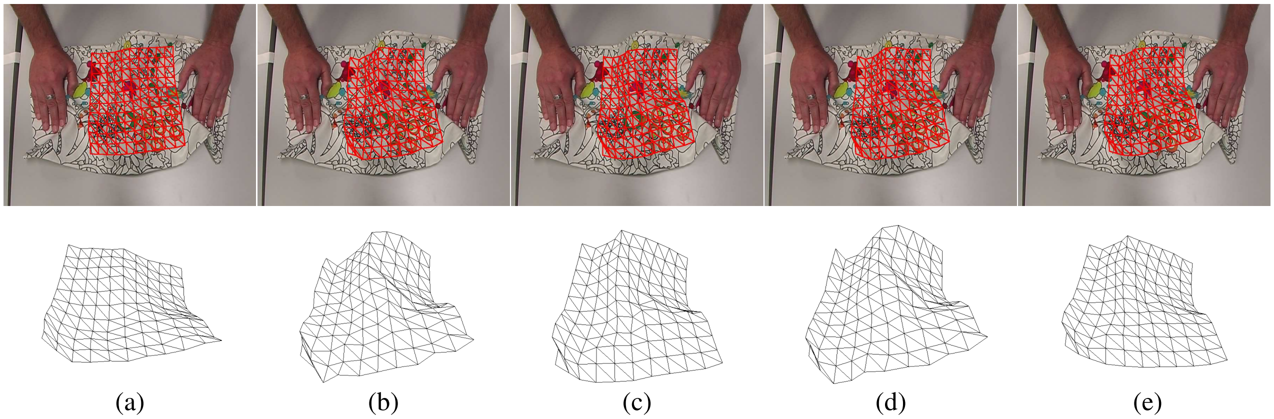


Figure 8. **Reconstructing a deforming piece of cloth.** Top row: Recovered mesh reprojected on the input image. Bottom row: Side view of the same mesh. Results obtained with (a) the original GP, the GP used with an image likelihood and constrained with (b) distance equalities and (c) inequalities, (d,e) same as (b,c) but when optimizing $\mathbf{k}_*$.

inextensibility constraints as our method, but in a frame-to-frame tracking context. Since our approach does not exploit temporal information, we initialized each frame with the reference shape. The second approach relies on distance inequalities instead of equality constraints. No regularizers were used for the baseline since the correspondences are well-spread over the surface. Fig. 6 depicts the MSE as a function of the noise variance for both datasets. Note that our approach yields more accurate reconstructions.

To show that our framework lets us encode different constraints, we replaced our distance preservation constraints by the distance inequalities of [10], which are better suited to represent folding surfaces. We first applied these constraints to the cardboard dataset with 2D locations as input. Since for this dataset equality constraints are well adapted, both methods perform similarly. We then acquired an additional dataset for a more flexible piece of cloth with a motion capture system. In this case, the distance inequalities outperformed the equality constraints. Results for both datasets are shown in Fig. 7.

Finally, we applied our method to real images. Fig. 1 depicts the results of reconstructing a deforming piece of paper. As it deforms smoothly, we used the cardboard data to train the GP. Note that only the constrained combination of discriminative and generative approaches correctly approximates the true shape, whereas the GP reconstruction noticeably stretches, and the constrained one fits the image less accurately. As a second test case, we used images of a deforming piece of cloth. Since we did not have training data corresponding to that particular surface, we used the inextensible meshes dataset. To account for the flexibility of the surface, we also used inequality constraints. Fig. 8 compares results obtained with both types of constraints. As is best observed from the video, inequality constraints yield more stable results. [1]

4.3. Articulated Pose Estimation

We now present the results of our approach on the articulated pose datasets. Fig. 9(left) depicts the MSE as a function of the silhouette noise using 250 training examples for the Poser body dataset [1]. Note that solely constraining the GP prediction does not improve the results. This can be explained by the fact that, when a joint angle is predicted inaccurately, making the limbs of the correct length may move the end-points even further from the ground-truth. This was not the case for deformable surfaces where each point is constrained by several neighbors. However, using our constraints in conjunction with an inverse GP yields a significant improvement. In Fig. 9(right), we show the MSE as a function of the number of training examples for 3% of switched pixels. As expected, the improvement is larger for smaller training sets.

Fig. 10 depicts our results for the different features computed on the hand dataset. Note that GPs trained from histograms of SIFT and PHOG features already perform well. However, in the case of steerable filters, while the GP performs poorly, our approach successfully reduced the error to a value comparable to those of the other features, and yields meaningful solutions, as can be seen in Fig. 10(right).

Fig. 11(a) shows the results obtained using the different features computed on the HumanEva [15] walking images of [7]. As the subject walks in circles, we trained our method on one loop and tested it on the rest of the sequence. Since for this dataset, estimating the pose from a single image is known to be ambiguous, we introduced a constant speed motion model in the loss function of Eq. 10, and solve for 20 frames simultaneously. Fig. 11(b) depicts our results for multiple activities when using the 2D joint locations as inputs. Since doing so removes some ambiguities of the problem, no dynamics were required. Note that our constrained combination of discriminative and generative methods improves the pose estimation.

[1] The videos for both sequences are available on the conference proceedings DVD and from the first author's webpage.

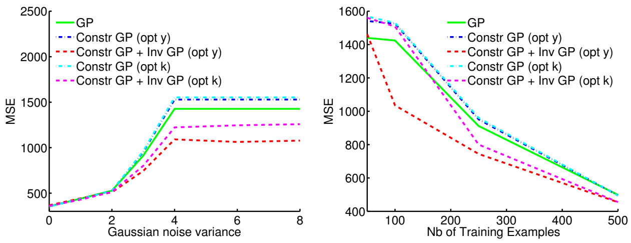


Figure 9. **Estimating human pose from silhouettes [1].** MSE of the original GP and of our method as a function of the percentage of noise in the silhouette (left), and as a function of the number of training examples (right).

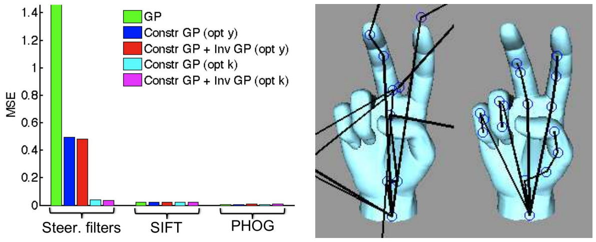


Figure 10. **Recovering the pose of a hand.** We compare the MSE obtained with several features. Note that the GP learned from steerable filters yields a large error that is reduced with our approach. Right: Even for a failure of the GP (non-typical), our method recovers the correct pose.

5. Conclusion and Future Work

In this paper, we have presented a unified framework for articulated pose estimation and non-rigid shape recovery. We have shown that introducing distance constraints into discriminative methods improved their performance on these tasks. However, these constraints are not sufficient since they do not prevent the recovered pose from drifting away from the pose depicted in the image. To overcome this problem, we have proposed a principled combination of discriminative and generative methods into a common formulation, which, we believe, is key to successfully address articulated and non-rigid pose estimation. In the future, we intend to study the use of different relaxations of our distance constraints, as well as different constraints, such as joint limits. We also plan to build more complex image likelihoods that rely on additional cues, such as shading.

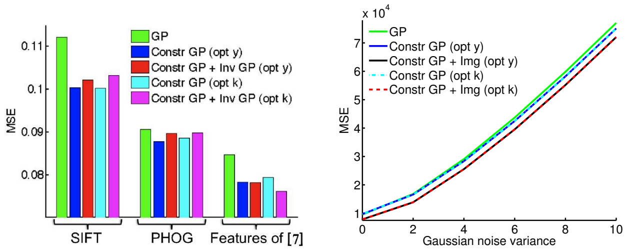


Figure 11. **Human pose estimation from real images [15].** (a) Results obtained from different features computed from the images of [7] and using a dynamical model. (b) Results obtained from 2D joint locations for several activities without a motion model. The different order of magnitudes of (a) and (b) are due to the use of different pose data.

References

[1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *CVPR*, 2004.

[2] J. Deutscher, A. Blake, and I. Reid. Articulated Body Motion Capture by Annealed Particle Filtering. In *CVPR*, 2000.

[3] D. J. C. MacKay. Introduction to Gaussian processes. In *Neural Networks and Machine Learning*, 1998.

[4] G. Mori and J. Malik. Recovering 3d human body configurations using shape contexts. *PAMI*, 2006.

[5] M. Perriollat, R. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. In *BMVC*, 2008.

[6] J. Pilet, V. Lepetit, and P. Fua. Real-Time Non-Rigid Surface Detection. In *CVPR*, 2005.

[7] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. Torr. Randomized Trees for Human Pose Detection. In *CVPR*, 2008.

[8] R. Rosales and S. Sclaroff. Learning Body Pose via Specialized Maps. In *NIPS*, 2002.

[9] R. Rosales and S. Sclaroff. Combining Generative and Discriminative Models in a Framework for Articulated Pose Estimation. In *IJCV*, 2006.

[10] M. Salzmann and P. Fua. Reconstructing sharply folding surfaces: A convex formulation. In *CVPR*, 2009.

[11] M. Salzmann, J. Pilet, S. Ilić, and P. Fua. Surface Deformation Models for Non-Rigid 3–D Shape Recovery. *PAMI*, 2007.

[12] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *COLT*, 2001.

[13] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *ICCV*, 2003.

[14] S. Shen, W. Shi, and Y. Liu Monocular template-based tracking of inextensible deformable surfaces under l2-norm. In *ACCV*, 2009.

[15] L. Sigal and M. J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report, 2006.

[16] L. Sigal, A. O. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*, 2007.

[17] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative Density Propagation for 3–D Human Motion Estimation. In *CVPR*, 2005.

[18] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Learning Joint Top-down and Bottom-up Processes for 3D Visual Inference. In *CVPR*, 2006.

[19] R. Urtasun and T. Darrell. Sparse Probabilistic Regression for Activity-independent Human Pose Inference. In *CVPR*, 2008.

[20] R. Urtasun, D. Fleet, and P. Fua. 3D People Tracking with Gaussian Process Dynamical Models. In *CVPR*, 2006.

[21] J. Zhu, S. C. Hoi, and M. R. Lyu. Nonrigid shape recovery by gaussian process regression. In *CVPR*, 2009.

[22] J. Zhu, S. C. Hoi, Z. Xu, and M. R. Lyu. An effective approach to 3d deformable surface tracking. In *ECCV*, 2008.