A Constrained Combination of Discriminative and Generative Methods

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

Raquel Urtasun
TTI - Chicago
rurtasun@ttic.edu

1. Introduction

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 image correspondences [2, 3, 6]. In contrast, many techniques for pose estimation have followed a discriminative approach, which allows for the use of more general image cues [1, 4, 5]. However, these techniques typically require large training sets and suffer from the fact that they assume the output dimensions to be independent given the inputs.

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). 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.

2. Approach

More formally, let $\hat{\mathbf{f}}$ be the estimate of the possibly non-linear mapping $\mathbf{f} : \Re^Q \rightarrow \Re^D$, such that $\mathbf{y} = \mathbf{f}(\mathbf{x}) + \epsilon$, learned by empirical risk minimization. Typically, when $\mathbf{y}$ is multi-dimensional, its dimensions are assumed to be independent given the inputs [1, 5]. As a consequence, the prediction $\hat{\mathbf{f}}(\mathbf{x}_*)$ for a new input $\mathbf{x}_*$ might not satisfy the constraints between the output dimensions.

For our particular case, let $\mathbf{y} \in \Re^{3N_p}$ be the vector of 3D coordinates of the N_p points that define a pose. Let $\mathcal{E}$ be the set of N_e links between 3D points whose length should remain constant. Finding a pose that satisfies those constraints can be formulated as solving the problem

$$\begin{aligned} &\underset{\mathbf{y}}{\text{minimize}} \;\; \mathcal{L}(\cdot, \mathbf{y}) + \lambda ||\hat{\mathbf{f}}(\mathbf{x}_*) - \mathbf{y}||_2^2 \\ &\text{subject to} \;\; ||\mathbf{y}_k - \mathbf{y}_j||_2^2 = l_{j,k}^2 \;,\; \forall (j,k) \in \mathcal{E} \;, \end{aligned} \tag{1}$$

where $\mathcal{L}(\cdot, \mathbf{y})$ is a loss function that depends on the image, $\mathbf{y}_i$ 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.

Since the constraints are non-convex, we propose to iteratively linearize them and solve the resulting problem. Let $\mathbf{y}_t$ be the estimate of the 3D shape at iteration t. We can approximate our constraints with their first order Taylor expansion, and find the displacement $\delta\mathbf{y}_t$ that satisfies the linearized constraints by solving $\mathbf{J}_t \delta\mathbf{y}_t = \mathbf{g}_t$, where $\mathbf{J}_t$ is the constraints Jacobian matrix evaluated in $\mathbf{y}_t$ and $\mathbf{g}_t$ contains the constraints errors for $\mathbf{y}_t$. Since this system has more unknowns than equations, it defines the family of poses

$$\mathbf{s}(\boldsymbol{\gamma}_t) = \mathbf{y}_t + \mathbf{J}_t^+ \mathbf{g}_t + \mathbf{V}_t^T \boldsymbol{\gamma}_t \;, \tag{2}$$

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

To encourage a stronger dependency on the predictor, we can rely on the *Representer theorem*. Doing so lets us define the prediction $\mathbf{y} = \boldsymbol{\alpha}\mathbf{k}_*$, where $\boldsymbol{\alpha}$ is learned from the training examples, and where we treat $\mathbf{k}_*$ as the new unknowns of our problem. Since $\mathbf{y}$ is a linear function of $\mathbf{k}_*$, we can follow a similar approach as before to encourage the predicted pose to satisfy constraints. In this case, this yields a family of poses defined as

$$\mathbf{s}(\boldsymbol{\gamma}_t) = \boldsymbol{\alpha} \cdot \left(\mathbf{k}_{*,t} + \mathbf{J}_t^+ \mathbf{g}_t + \mathbf{V}_t^T \boldsymbol{\gamma}_t\right) \;. \tag{3}$$

Given the new unknowns $\boldsymbol{\gamma}_t$ that implicitly minimize the violation of the constraints, we can re-write Eq. 1 as

$$\underset{\boldsymbol{\gamma}_t}{\text{minimize}}\, \mathcal{L}(\cdot, \boldsymbol{\gamma}_t) + \lambda ||\hat{f}(\mathbf{x}_*) - \mathbf{s}(\boldsymbol{\gamma}_t)||_2^2 \;, \tag{4}$$

where $\mathbf{s}(\boldsymbol{\gamma}_t)$ is given either by Eq. 2 or by Eq. 3. For a squared loss function $\mathcal{L}(\cdot, \boldsymbol{\gamma}_t)$, this is a convex optimization problem, whose minimum can be obtained in closed-form by solving a linear system in the least-squares sense.

3. Experimental evaluation

In practice, we used Gaussian processes as our discriminative predictor. Furthermore, the image-based loss function relied either on an inverse mapping from $\mathbf{y}$ to $\mathbf{x}$, on the reprojection error of 3D points, or on template matching and boundary information.

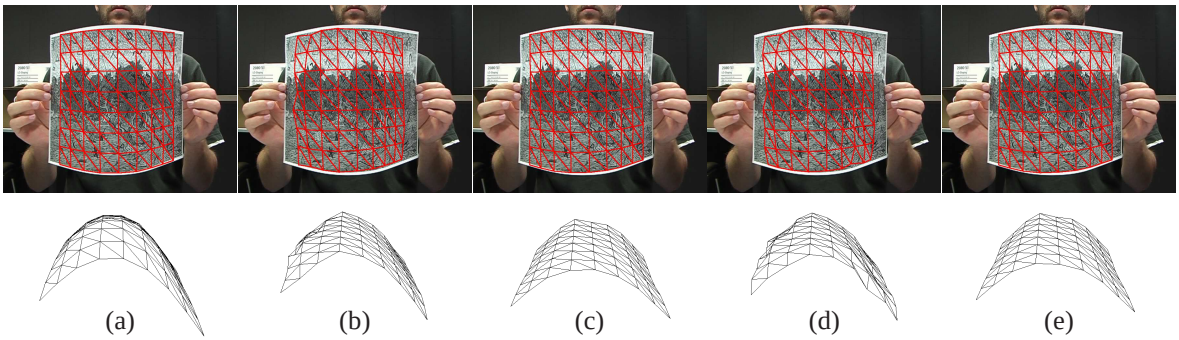

Figure 1. **Reconstructing a piece of paper from monocular images.** Top row: Recovered mesh reprojected on the input image. Bottom row: Side view of the same mesh. Results were obtained with (a) the original predictor, (b) the constrained predictor, (c) the constrained predictor with an image likelihood, (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.

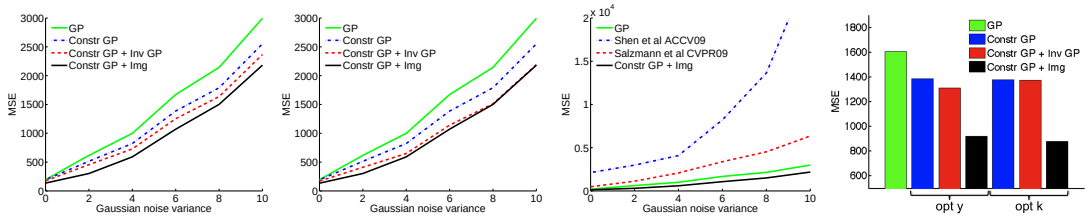

Figure 2. **Reconstructing a deforming piece of cardboard.** (a)-(c) Average MSE as a function of the input noise variance by optimizing (a,c) $\mathbf{y}$ and (b) $\mathbf{k}_*$. (c) Comparison against [3] and [2]. The inputs are taken as (a)-(c) the image locations of the mesh vertices, and (d) spatial pyramid of HOG features.

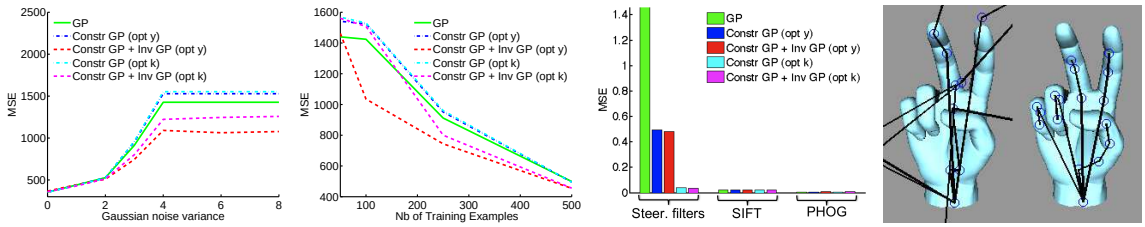

Figure 3. **Estimating articulated pose** (a,b) Human pose from silhouettes [1]. MSE of the original GP and of our method as a function of the percentage of noise in the silhouette (a), and as a function of the number of training examples (b). (c,d) Hand pose from image features. (d) Even for a failure of the GP (non-typical), our method recovers the correct pose.

Fig. 1 depicts the results of different approaches when reconstructing a piece of paper from real monocular images. The image locations of the mesh vertices were taken as inputs, and we minimized the vertices reprojection error.

Fig. 2 depicts results obtained when reconstructing a piece of cardboard from synthetic data. Fig. 2 (a,b) show the mean reconstruction error (MSE) as a function of the input noise variance for 250 training examples. In Fig. 2 (c), we compare our approach against [3, 2]. The first method relies on the same constraints as ours, 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. Fig. 2 (d) depicts the MSE obtained from PHOG features.

In Fig. 3, we show results on articulated pose estimation. For human pose estimation, Fig. 3 (a,b) depicts the MSE of the original GP and of our method as a function of the percentage of noise in the silhouette (a), and as a function of the number of training examples (b). Fig. 3 (c,d) shows our results for hand pose estimation. Note that, even when the GP fails, our method recovers the correct pose.

References

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

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

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

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

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

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