Hierarchical Data-driven Descent for Efficient Optimal Deformation Estimation

Yuandong Tian
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
yuandong@cs.cmu.edu

Srinivasa G. Narasimhan
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
srinivas@cs.cmu.edu

Abstract

Real-world surfaces such as clothing, water and human body deform in complex ways. The image distortions observed are high-dimensional and non-linear, making it hard to estimate these deformations accurately. The recent data-driven descent approach [17] applies Nearest Neighbor estimators iteratively on a particular distribution of training samples to obtain a globally optimal and dense deformation field between a template and a distorted image. In this work, we develop a hierarchical structure for the Nearest Neighbor estimators, each of which can have only a local image support. We demonstrate in both theory and practice that this algorithm has several advantages over the non-hierarchical version: it guarantees global optimality with significantly fewer training samples, is several orders faster, provides a metric to decide whether a given image is "hard" (or "easy") requiring more (or less) samples, and can handle more complex scenes that include both global motion and local deformation. The proposed algorithm successfully tracks a broad range of non-rigid scenes including water, clothing, and medical images, and compares favorably against several other deformation estimation and tracking approaches that do not provide optimality guarantees.

1. Introduction

Accurately finding dense correspondence between images capturing deforming objects is important for many vision tasks, such as 3D reconstruction, image alignment and tracking. However, estimating the parameters of non-rigid deformation is hard due to its high-dimensionality and strong nonconvexity. Continuous optimization approaches (e.g. gradient descent or Newton's method) require no training but often suffer from local minima, while regression-based approaches (e.g., Nearest Neighbor) have guaranteed solutions, but need a lot of training samples.

Recently, Tian and Narasimhan [17] proposed *Data-driven Descent* which combines the best properties of both continuous optimization and regression. They show that if a generative model for deformation is available, then the training samples can be generated by simply deforming the template using parameters from a particular distribution. Then a sequence of Nearest Neighbor predictions will achieve the globally optimal solution that warps the test image to the template. Furthermore, to achieve the accuracy of $1/\epsilon$, the number of samples needed is $O(C^d \log 1/\epsilon)$ for d dimensional warping, much less than $O(1/\epsilon^d)$ required for general regressions. Intuitively, this approach captures the group-like structure in deformation and uses the training samples which are far away from the test image for prediction. Their approach shows good empirical results for local deformation, but fails to capture general deformation that contains both global and local components (e.g., cloth moving and deforming).

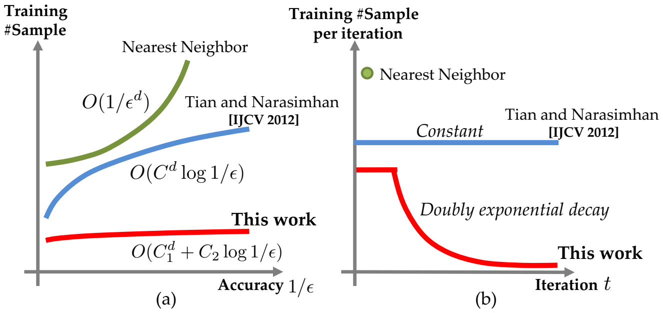


Figure 1. Illustrations of order of training sample complexity required for estimating d dimensional deformation. **(a):** To achieve a guaranteed accuracy $1/\epsilon$, traditional regression-based approaches (e.g. Nearest Neighbor) require $O(1/\epsilon^d)$ training samples. Data-driven Descent [17] requires $O(C^d \log 1/\epsilon)$, decoupling the dimensionality from the accuracy. Our hierarchical framework for deformation estimation achieves $O(C_1^d + C_2 \log 1/\epsilon)$ with constant C_1 much smaller than C and C_2 independent of dimensionality. **(b)** Sample complexity per iteration. A constant number of samples per iteration is needed in [17]. The number of samples needed is a constant for the first few iterations, and then decays *double exponentially* for our algorithm.

In this paper, we develop a top-down hierarchical structure for deformation estimation with global optimality guarantee. First, the deformation field is parameterized so that the deformation happening within a local image patch can be predicted by the content of that patch, reducing the dimensionality. Then, we model the relationship between the image content and the deformation parameters using a novel criterion. With this criterion, all patches at different locations and scales can be regarded as predictors with guaranteed worst-case precisions. Finally, combining these predictors together in a top-down hierarchical manner leads to an overall predictor that can handle large and high-dimensional deformation with both local and global components.

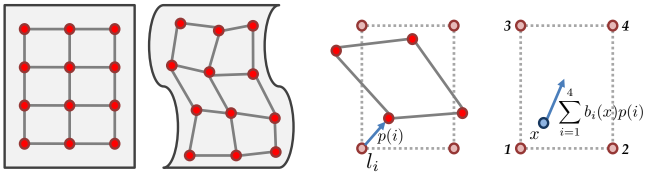
Our contributions are three-fold. *First*, our approach brings down sample complexity to $O(C_1^d + C_2 \log 1/\epsilon)$, which varies very slowly with respect to the accuracy. In particular, the number of samples required in each iteration stays constant for the first few iterations (layers of hierarchy), and then decays double exponentially (Fig. 1). Practically, our unoptimized Matlab implementation is fast, achieving 3-4 fps on real images. *Second*, compared to [17], our sample complexity guarantee is based on much weaker assumptions that can be verified with an efficient algorithm. As a result, our constant C_1 is much smaller than the constant C in [17]. *Third*, our work provides a rigorous theoretical analysis and interesting insights for top-down coarse-to-fine hierarchical structures. We believe this can be useful for analyzing many other hierarchies proposed in the computer vision community.

Our work not only has strong theoretical foundations, but also demonstrates good quantitative and qualitative results on real video sequences containing different types of deformation, including clothing and water surface deformations as well as medical images of internal organs. Our approach outperforms optimization-based approaches such as Lucas-Kanade [1] and Free-form registration [9] (both with coarse-to-fine implementations), regression-based approaches such as Nearest Neighbor and Explicit Shape Regression [3], feature-based approaches such as SIFT [6], tracking-based approaches such as KLT [13], and hybrid methods such as Data-driven Descent [17].

2. Related Works

Optimization-based approaches (e.g., [1, 7, 9]) usually reach a local minimum using gradient descent or Newton's method. Random initialization is used to improve the quality of solutions on a heuristic basis. Regression-based approaches aim to learn a mapping from the distorted image to the deformation parameters using labeled training samples. The actual form of mapping could be Nearest Neighbor, Linear [7], Random Forest [14], Boosted Random Fern [3], etc. Feature-based approaches (e.g., SIFT [6]) find correspondence by matching local features, and have to balance between distinctiveness and invariance under deformation.

Hierarchical structures have been used extensively in vision. Typical scenarios include coarse-to-fine optimization [9] for a better local solution, interest point detection [6], multi-resolutional feature extraction [5], biologically plausible framework for object recognition [12] and so on. Recently, it is also used in Deep Learning, showing state-of-art performance in image classification [4]. However, as far as we know, none of the previous works provide theoretical performance guarantees.



(a) Template T (b) Distorted Image I_p (c) Displacement $p(i)$ for landmark l_i (d) Displacement at pixel x

Figure 2. Local parameterization of deformation. **(a)-(b)** The deformation field is controlled by a set of landmarks on the template image. By moving these landmarks, a deformed image is created. **(c)** Local parameterization. Each parameter $\mathbf{p}(i)$ encodes the 2D displacement of the landmark i. **(d)** Displacement on any pixel $\mathbf{x}$ is interpolated using displacements of nearby landmarks.

3. The Image Deformation Model

Denote T as the template image and $I_\mathbf{p}$ as the distorted image with deformation parameters $\mathbf{p}$. The deformation field $W(\mathbf{x};\mathbf{p})$ maps the pixel location $\mathbf{x}$ on the template to the pixel location $W(\mathbf{x};\mathbf{p})$ on the distorted image $I_\mathbf{p}$:

$$I_\mathbf{p}(W(\mathbf{x};\mathbf{p})) = T(\mathbf{x}) \tag{1}$$

We *locally parameterize* the deformation field $W(\mathbf{x};\mathbf{p})$ at any 2D point $\mathbf{x}$ by a weighted linear combination of displacements $\mathbf{p} = [\mathbf{p}(1), \mathbf{p}(2), \ldots, \mathbf{p}(K)]^T$ on K landmarks (Fig. 2):

$$W(\mathbf{x};\mathbf{p}) = \mathbf{x} + B(\mathbf{x})\mathbf{p} \tag{2}$$

where $B(\mathbf{x}) = [b_1(\mathbf{x}), b_2(\mathbf{x}), \ldots, b_K(\mathbf{x})]$ is a K-dimensional row vector of weighting factors on location $\mathbf{x}$ from K landmarks, $\mathbf{p}$ is a K-by-2 matrix storing $2K$ displacement components and $\mathbf{p}(i)$ is a 2-dimensional deformation vector for landmark i. Due to strong correlations between nearby landmark displacements, the dimensionality d of the warping field could be much lower than $2K$.

Naturally, at any location $\mathbf{x}$, its weights from all K landmarks add to 1 ($\sum_i b_i(\mathbf{x}) = 1$), and the weight $b_i(\cdot)$ at i-th landmark location l_i is 1 while others are zero. Practically, $B(\mathbf{x})$ can be any interpolation function, e.g., Thin-plate Spline [2], B-spline [9], local linear interpolation, etc.

Many previous works [17, 7, 11] also assume similar form of $W(\mathbf{x};\mathbf{p})$. However, their parameters $\mathbf{p}$, usually given by dimensionality reduction procedures (e.g., PCA), is not localized to spatial landmarks, while Eqn. 2 is an *over-parameterization* of the deformation field $W(\mathbf{x};\mathbf{p})$, which leads to further reduction of training samples needed.

Generating training samples. From Eqn. 1, given the parameter $\mathbf{p}$, one can generate the deformed image $I_\mathbf{p}$ from the template T. This is done by assigning every pixel $\mathbf{y}$ of the deformed image $I_\mathbf{p}$ with the pixel value on location $\mathbf{x} = W^{-1}(\mathbf{y};\mathbf{p})$ of the template T. Choosing different parameters $\{\mathbf{p}_i\}$ gives many *training samples* $\{(\mathbf{p}_i, I_{\mathbf{p}_i})\}$. The task now becomes how to properly distribute the training samples and how many samples are needed (i.e., *sam-*

ple complexity) to achieve the globally optimal prediction of the unknown parameter for a distorted test image. This is the core of our contribution and will be described next.

4. The Relationship between Image Evidence and Distortion Parameters

Suppose we have training samples $\{\mathbf{p}, I_{\mathbf{p}}\}$ and want to predict the parameter for a test image I with an unknown true parameter $\mathbf{p}_1$. The simplest way is to use the Nearest Neighbor predictor: find $I_{\mathbf{p}_2}$ in the training set that is closest to I, and return the parameter $\mathbf{p}_2$ as the prediction.

To make this approach work, we need to assume a positive correlation between the image difference $\Delta I \equiv \|I_{\mathbf{p}_1} - I_{\mathbf{p}_2}\|$ in terms of a certain image metric and the parameter difference $\Delta\mathbf{p} \equiv \|\mathbf{p}_1 - \mathbf{p}_2\|_\infty$ in terms of maximal absolute difference between landmark displacements. Intuitively, this means that if two images are close, so are their parameters and vice versa. This can be represented by the following Lipchitz conditions proposed in [17]:

$$L_1 \Delta I \leq \Delta\mathbf{p} \leq L_2 \Delta I \tag{3}$$

where, L_1 and L_2 are two constants that are dependent on the template T. [17] shows that the ratio L_2/L_1 is a charactistic for samples complexity for guaranteed Nearest Neighbor prediction. For simple images that contain one salient object with a clear background, L_2/L_1 is small and a few samples suffice. For difficult images with repetitive patterns, L_2/L_1 is large and a lot of samples are needed to distinguish among locally similar-looking structures.

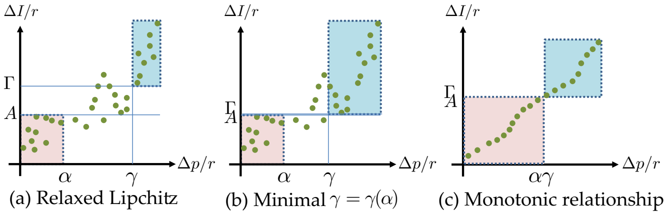
4.1. Relaxed Lipchitz Condition

One shortcoming of Eqn. 3 is that it must hold for arbitrarily small ΔI and $\Delta\mathbf{p}$. Thus it fails in two situations:

- **Noisy images.** Adding noise to a distorted image $I_{\mathbf{p}}$ changes its appearance but not its parameters. As a result, $\Delta\mathbf{p} \approx 0$ but ΔI is finite. This makes $L_1 \to 0$.
- **Repetitive Patterns.** If an image resembles itself after some transformation, then $\Delta\mathbf{p}$ is finite but $\Delta I \approx 0$. This makes $L_2 \to +\infty$.

In both cases, [17] gives a trivial (infinite) bound on sample complexity and global optimality cannot be guaranteed.

In this paper, we relax this condition using a patch-based approach. Denote $R = R(\mathbf{x}, r)$ as a square centered at x with size $2r$. $I(R)$ is the patch within R and $S = S(\mathbf{x}, r)$ is the subset of landmarks whose displacements $\mathbf{p}(S)$ influence the patch content $I(R)$. $\mathbf{p}(S)$ is a $|S|$ by 2 matrix obtained by choosing S rows from p. We assume $I(R)$ and $\mathbf{p}(S)$ satisfy the following *relaxed Lipchitz Condition*:



(a) Relaxed Lipchitz (b) Minimal $\gamma = \gamma(\alpha)$ (c) Monotonic relationship

Figure 3. Relaxed Lipchitz Condition (Eqn. 4). **(a)** Four constants $(\alpha, \gamma, A, \Gamma)$ capture the correlations between ΔI and $\Delta\mathbf{p}$. When $\Delta\mathbf{p}$ is small ($\leq \alpha r$), ΔI is small as well ($\leq Ar$). Conversely, with large parameter difference ($\geq \gamma r$), the image difference is also large ($\geq \Gamma r$). **(b)** Given α, there exists a minimal γ. **(c)** For a monotonic relationship between ΔI and $\Delta\mathbf{p}$, $\alpha = \gamma \in [0, 1]$.

Assumption 1 (Relaxed Lipchitz Condition) *There exists $0 < \alpha(\mathbf{x}, r) \leq \gamma(\mathbf{x}, r) < 1$ and $0 < A(\mathbf{x}, r) < \Gamma(\mathbf{x}, r)$ so that for any $\mathbf{p}_1$ and $\mathbf{p}_2$ with $\|\mathbf{p}_1\|_\infty \leq r$, $\|\mathbf{p}_2\|_\infty \leq r$:*

$$\Delta\mathbf{p} \leq \alpha r \implies \Delta I \leq Ar,\ \Delta\mathbf{p} \geq \gamma r \implies \Delta I \geq \Gamma r \tag{4}$$

for $\Delta\mathbf{p} \equiv \|\mathbf{p}_1(S) - \mathbf{p}_2(S)\|_\infty$ and $\Delta I \equiv \|I_{\mathbf{p}_1}(R) - I_{\mathbf{p}_2}(R)\|$.

Here $\|\mathbf{x}\|_\infty \equiv \max_i |x_i|$. Visually, the first part of Eqn. 4 says all $(\Delta\mathbf{p}, \Delta I)$ left to the vertical line αr have to be in the red-shaded box; while the second part says all $(\Delta\mathbf{p}, \Delta I)$ right to the vertical line γr have to be in the blue-shaded box (Fig. 3(a)). The condition $A < \Gamma$ means that the bottom of blue is always above the top of red. When they touch (Fig. 3(b)), the *minimal*, or tightest γ is achieved for a given α, which is the monotonous curve $\gamma = \gamma(\alpha)$.

Different from the Lipchitz conditions (Eqn. 3), one important aspect of Eqn. 4 is that ΔI and $\Delta\mathbf{p}$ are only correlated up to the scale of r. This *weaker* condition allows Eqn. 4 to account for noise and parameter changes outside the subset S that may influence the patch $I(R)$ without altering $\mathbf{p}(S)$. This also accounts for the case in which two slightly different parameters share the same image appearance. In both cases, the image-dependent pair (α, γ) is still well-behaved while L_2/L_1 is not. Another aspect is that Eqn. 4 is assumed for *every* patch located at $\mathbf{x}$ with scale r, while Eqn. 3 is a single condition for the entire image.

Besides, Eqn. 4 holds only for deformation within the *acceptance range* r, i.e., $\|\mathbf{p}\|_\infty \leq r$). This is a practical condition because if $\mathbf{p}(S)$ is large, the image content $I(R)$ is no longer related to the local patch deformation $\mathbf{p}(S)$.

Degrees of Freedom on patches. Since $\mathbf{p}(S)$ is a $|S|$-by-2 matrix, there are at most $2|S|$ apparent degrees of freedom for patch $I(R)$. How large is $|S|$? If landmarks are distributed uniformly (e.g., on a regular grid), $|S|$ is proportional to $Area(R)$, which gives $2|S| \propto r^2$.

On the other hand, if the overall effective degree of freedom is d, then no matter how large $2|S|$ is, $\mathbf{p}(S)$ contains dependent displacements and the effective degree of freedom in R never exceeds d. Therefore, we assume:

Assumption 2 (Degrees of Freedom for Patches) *The local degrees of freedom of a patch $(\mathbf{x}, r)$ is $\min(d, 2|S|)$.*

4.2. Guaranteed Prediction using Nearest Neighbor

Now let us study how the relaxed Lipchitz condition helps Nearest Neighbor prediction. We wish to know how well patch $(\mathbf{x}, r)$ can predict the deformation $\mathbf{p}(S)$ within its acceptance range r (i.e., $\|\mathbf{p}\|_\infty \le r$). Without any training samples, we can trivially set the prediction $\hat{\mathbf{p}}(S) = 0$ and get a worst-case guaranteed prediction error of r. Now the problem is: if we want to obtain a slightly better prediction, how many training samples do we need?

Theorem 1 gives the answer. It shows that if the relaxed Lipchitz condition (Eqn. 4) holds, then a Nearest Neighbor prediction with $1/\alpha$ samples per dimension will always reduce the error by a factor of $\gamma < 1$:

Theorem 1 (Guaranteed Nearest Neighbor) *Given a distorted image $I_\mathbf{p}$ with $\|\mathbf{p}\|_\infty \le r$, then with*

$$N(\mathbf{x}, r) = \min\left(c_{ss}\left\lceil 1/\alpha \right\rceil^d, \left\lceil 1/\alpha \right\rceil^{2|S|}\right) \qquad (5)$$

number of samples uniformly distributed in the hypercube $[-r, r]^{2|S|}$, we can compute a prediction $\hat{\mathbf{p}}(S)$ so that

$$\|\hat{\mathbf{p}}(S) - \mathbf{p}(S)\| \le \gamma r \qquad (6)$$

using Nearest Neighbor in the region R with image metric.

Proof Sketch We first fill the $2|S|$-dimensional hypercube $[-r, r]^{2|S|}$ with $(1/\alpha)^{2|S|}$ training samples uniformly. Then, for any test sample I within, there is I' whose parameter difference is within αr. By Eqn. 4, $\|I - I'\| \le Ar$. The nearest neighbor of I, namely I_{NN}, is closer to I than I' to I. Again by Eqn. 4, the parameter of I_{NN}, which is the prediction, is γr close to the true parameters of I.

If the local deformation is d-dimensional with $d < 2|S|$, then it turns out that only a small fraction of the hypercube are sampled and $c_{ss}\left\lceil 1/\alpha \right\rceil^d$ samples suffices. See [16] for detailed derivation of c_{ss}. ■

From Theorem 1, the exponent of Eqn. 5 is the degrees of freedom mentioned in Assumption 2, which demonstrates the curse of dimensionality. From Eqn. 5 and Eqn. 6, now both α and γ have their physical meanings: α is the inverse of sample complexity per dimension, while γ is the inverse of prediction accuracy. Ideally we want α to be large for lower sample complexity, and γ to be small for higher accuracy. However, the constraint $\alpha \le \gamma$ and the minimal curve $\gamma = \gamma(\alpha)$ show there is a trade-off. Like L_2/L_1 in Eqn. 3, this trade-off reflects the difficulty level of images for deformation prediction (See Sec. 6 for details).

5. Construction of Hierarchical Structure

According to Theorem 1, different image patches $(\mathbf{x}, r)$ show different characteristics in their prediction guarantees: large patches (large r) can deal with large deformation but have low prediction precision, while small patches (small r) only deals with small deformation but enjoys high prediction precision. Therefore, in order to estimate large deformation with high precision, a natural way is to build a coarse-to-fine hierarchy of predictions as follows: the coarse layer (large patch) reduces the prediction residue by a certain extent so that it is within the acceptance range of the fine layer (small patch), where the prediction is refined.

From this argument, we construct the hierarchical structure as follows. Within the same layer t, scale of patches is fixed and denoted as r_t. When going from top to bottom (t becomes large), the scale r_t of patches shrinks towards zero. The shrinking factor $\bar{\gamma} = r_{t+1}/r_t$ is set to be

$$\bar{\gamma} \equiv \max_{(\mathbf{x}, r)} \gamma(\mathbf{x}, r) < 1 \qquad (7)$$

Algorithm 1 Hierarchical Deformation Estimation.

```
INPUT Training samples $Tr(\mathbf{x}, r) \equiv \{(\mathbf{p}_i, I_i)\}$ for each image patch $(\mathbf{x}, r)$.
INPUT Test image $I_{\text{test}}$ with unknown parameters $\mathbf{p}$.
Set an initial estimation $\hat{\mathbf{p}}^0 = 0$.
for $t = 1$ to $T$ do
    Set the current image $I_c(\mathbf{x}) = I_{\text{test}}(W(\mathbf{x}; \hat{\mathbf{p}}^{t-1}))$.
    for Patch $(\mathbf{x}_j, r_t)$ within layer $t$ do
        $S_j = S(\mathbf{x}_j, r_t)$, $R_j = R(\mathbf{x}_j, r_t)$
        Find the Nearest Neighbor $i^*$ for patch $I(R_j)$:
        $i^* = \arg\min_{i \in Tr(\mathbf{x}, r)} \|I_c(R_j) - I_i(R_j)\|$
        Set the estimation $\tilde{\mathbf{p}}_{j \to i}(S_j) = \mathbf{p}_{i^*}(S_j)$.
    end for
    Aggregation: $\tilde{\mathbf{p}}(i) = \text{mean}_{i \in S_j} \tilde{\mathbf{p}}_{j \to i}(S_j)$.
    Update: $\hat{\mathbf{p}}^t(i) = \hat{\mathbf{p}}^{t-1}(i) + \tilde{\mathbf{p}}(i)$ for all landmarks.
end for
Return final predictions $\hat{\mathbf{p}}^T(i)$ for all landmarks.
```

Fig. 4 and Alg. 1 illustrate the algorithm that estimates the unknown parameter p given the test image I_{test}. For the first iteration, the test image I_{test} is directly compared with the training samples generated from the entire image with scale r_1 to obtain the Nearest Neighbor prediction $\hat{\mathbf{p}}^1$. Then for the second iteration, we have a slightly less distorted image $I_{\text{test}}(W(\mathbf{x}, \hat{\mathbf{p}}^1))$, from which we estimate $\mathbf{p} - \hat{\mathbf{p}}^1$. Since $\|\mathbf{p} - \hat{\mathbf{p}}^1\|_\infty$ is smaller than $\|\mathbf{p}\|_\infty$, its predictions can be localized to smaller patches. Then this procedure is iterated until the lowest layer. Similar to [17], Alg. 1 will converge to the globally optimal solution (Theorem 2), while the required number of samples is $O(C_1^d + C_2 \log 1/\epsilon)$ (Theorem 3).

Note that a less distorted image $I_{\text{test}}(W(\mathbf{x}; \hat{\mathbf{p}}^{t-1}))$, as the input of layer t, is not necessarily the same as a distorted image $I_{\mathbf{p} - \hat{\mathbf{p}}^{t-1}}$ generated directly from the template image. However, their difference decreases when $r_t \to 0$ and global convergence guarantee still holds (See [16]).

Theorem 2 (The Global Convergence Theorem) *If $\|\mathbf{p}\|_\infty \le r_1$, then the prediction $\hat{\mathbf{p}}^t(i)$ satisfies:*

$$\|\hat{\mathbf{p}}^t(i) - \mathbf{p}(i)\|_\infty \le \bar{\gamma}^t r_1 \to 0 \quad \text{when} \quad t \to +\infty \qquad (8)$$

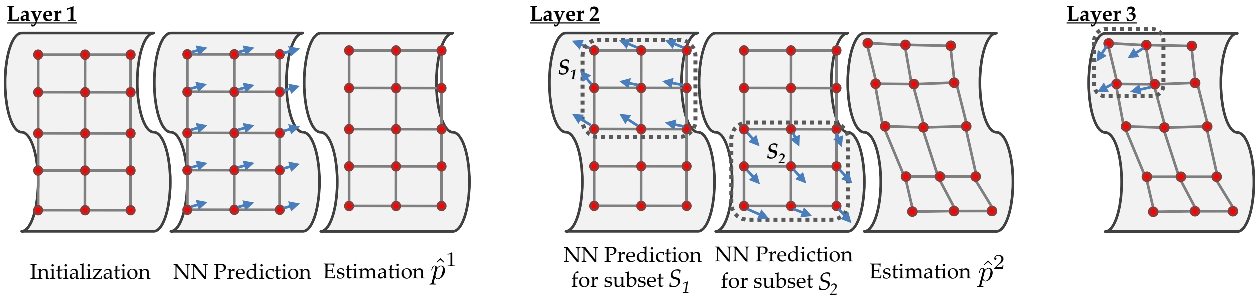
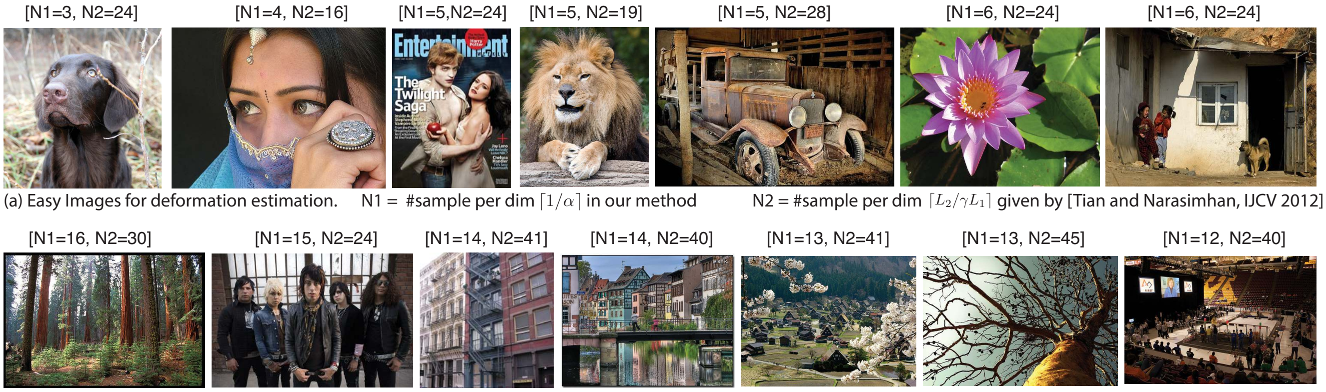


Figure 4. Work flow of our hierarchical algorithm for deformation estimation. On Layer 1, a global prediction is made and the estimation is updated. On Layer 2, local deformation is estimated and aggregated. The procedure repeats until the last layer.



(a) Easy Images for deformation estimation. N1 = #sample per dim $\lceil 1/\alpha \rceil$ in our method N2 = #sample per dim $\lceil L_2/\gamma L_1 \rceil$ given by [Tian and Narasimhan, IJCV 2012]

(b) Hard Images for deformation estimation.

Figure 5. Exemplar images and the theoretical bounds for the number of samples needed per dimension. For each bracket, the first number is our bound (given by $1/\alpha_{0.95}$), while the second number from Data-Driven Descent (given by $L_2/\gamma L_1$ with $\gamma = 0.95$). **Top Row:** Images with a salient object and clean background require only a few samples per dimension. **Bottom Row:** Images with repetitive patterns require more samples per dimension. In both cases, our bound is smaller than that given by Data-Driven Descent.

Proof Sketch From Theorem 1, from the top layer, after each layer the residue is contracted by at least $\bar{\gamma}$. Then $\bar{\gamma} < 1$ implies that the error diminishes from top to bottom. ∎

Theorem 3 (The Number of Samples Needed) *The total number N of samples needed is bounded by:*

$$N \leq C_3 C_1^d + C_2 \log_{1/\bar{\gamma}} 1/\epsilon \tag{9}$$

where $C_1 = 1/\min_{(\mathbf{x},r)} \alpha(\mathbf{x}, r)$, $C_2 = 2^{1/(1-\bar{\gamma}^2)}$ *and* $C_3 = 2 + c_{ss}(\lceil \frac{1}{2} \log_{1/\bar{\gamma}} 2K/d \rceil + 1)$.

Proof Sketch From Assumption 2, the areas of patches, as well as $|S|$, decrease by a factor of $\bar{\gamma}^2$ from top to bottom. Therefore, the number of samples needed stays the same until $2|S| \approx d$, and then goes down *double-exponentially*. Theorem 1 gives the number of samples at any level of the hierarchy. The supplementary report [16] shows that the summation of the samples at all levels is a fast decaying series bounded by Eqn. 9. ∎

6. Empirical Upper Bounds For Images

Given a specific template and a specific family of deformation, we can generate many deformed images and their parameters $(\mathbf{p}_i, I_{\mathbf{p}_i})$, compute all-pair image/parameter distances $\{\Delta \mathbf{p}_i, \Delta I_i\}$ and estimate the monotonous curve $\gamma = \gamma(\alpha)$ like Fig. 3. This curve can help predict the theoretical difficulties of images for deformation estimation. For simplicity, we set a constant and convergent contraction factor $\bar{\gamma} = 0.95$ and compute the largest $\alpha_{0.95} = \gamma^{-1}(0.95)$. Therefore, simple images have high $\alpha_{0.95}$, indicating low sample complexity per dimension $(1/\alpha_{0.95})$, and vice versa.

We randomly generate 1000 deformed samples and compute all-pair distances. The deformation is 2D translation and in-plane rotation ($d = 3$) up to $\pm\pi/8$. We propose Alg. 2 which only costs $O(M \log M)$ to compute the curve $\gamma = \gamma(\alpha)$, while brute-force search costs $O(M^3)$.

Fig. 5 shows each template and its $1/\alpha_{0.95}$. Note that images with a salient object and uniform background requires fewer samples, while images with repetitive patterns and cluttered backgrounds require more. In contrast, $L_2/\gamma L_1$, as suggested in [17], is much higher in both cases.

With regard to total sample complexity N, Theorem 1 tells that for easy images, $1/\alpha_{0.95} \approx 5$ and $N \approx [5 \cdot (2 + \sqrt{2})]^4 = 84926$ (See [16] for details), while for hard images, $1/\alpha_{0.95} \approx 12$ and $N \approx [12 \cdot (2 + \sqrt{2})]^4 = 2817654$. Although practically N may be much smaller, it gives a sense of difficulty levels of images.

Algorithm 2 Find Local Lipschitz Constants.

```
INPUT Parameter distances $\{\Delta \mathbf{p}_i\}$ with $\Delta \mathbf{p}_i \le \Delta \mathbf{p}_{i+1}$.
Image distances $\{\Delta I_i\}$ and scale $r$.
$\Delta I_i^+ = \max_{1 \le j \le i} \Delta I_j$, for $i = 1 \dots M$.
$\Delta I_i^- = \min_{i \le j \le M} \Delta I_j$, for $i = 1 \dots M$.
for $i = 1$ to $M$ do
  Find minimal $j$ so that $\Delta I_j^- > \Delta I_i^+$.
  if $i \le j$ then
    Store a curve point $(\alpha, \gamma) = (\Delta \mathbf{p}_i, \Delta \mathbf{p}_j)/r$.
  end if
end for
```

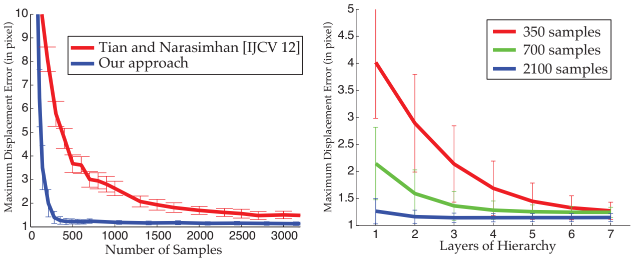


Figure 6. Performance of the proposed algorithm. **Left:** Performance comparison with [17]. Accuracy of our approach increases much faster than [17] with the same number of samples. To obtain the same level of accuracy of our approach with 400 samples, [17] requires 10000 samples or more. Our approach also has lower variance in performance. **Right:** Convergence behavior of our approach with different number of training samples.

7. Experiments on Simulated Data

Our algorithm works well for synthetic data. For all the experiments, our approach adopts a hierarchical structure using a grid of 256 landmarks with $\bar{\gamma} = 0.7$ and $T = 8$ layers. For bases functions, we use Thin-Plate Spline [2] proper normalization. While our theory gives an upper bound of the sample complexity, practically 350 training samples over all layers suffice for good performance.

7.1. Convergence Behavior

We artificially distorted 100 images with a 20-dimensional global warping field specified in [17]. For each image, its 10 distorted versions are generated with random parameters, which are estimated using Data-driven Descent (TN) [17] and using our approach.

Fig. 6 shows the performance comparison. Our algorithm obtains much better performance and lower variance compared to TN with the same number of training samples. Note that the strong drop in error shows that our method achieves very high accuracy by adding very few samples once it starts to work. This coincides with Fig. 1.

7.2. Deformation Estimation on Repetitive Patterns

We further test our approach on synthetic data containing distorted repetitive patterns, and compare it with previous methods. From an undistorted template (240-by-240), we generate a dataset of 200 distorted images, each with labeled 49 points. The deformation field is created by random Gaussian noise without temporal continuity.

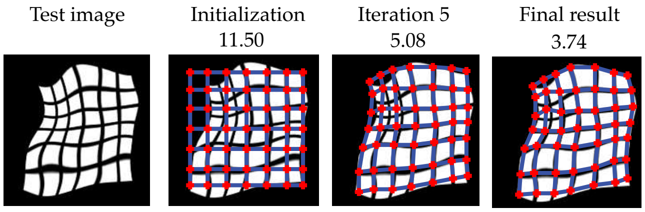


Figure 7. Demonstration of the iterative procedure of our algorithm. Starting from initialization, the algorithm applies predictors of different layers to estimate the landmark locations. Numbers on top show RMS errors.

The overall degree of freedom for this dataset is very high (50 dimensions are needed to achieve < 1 pixel reconstruction error). It is in general impossible to have sufficient number of samples for global optimality conditions to be satisfied. However, practically our method still works well.

We compare our approach to the following previous methods: Lucas-Kanade (LK) [1], Data-driven Descent (TN) [17], Free-form registration (FF) [9], Explicit Shape Regression (ESR) [3] and SIFT matching with outlier removal using RANSAC (SR) [6]. LK and TN use a local parametric deformation model. LK uses local affine bases of size 100-by-100, and TN uses a 20-dimensional smooth bases of size 57-by-40 [17]. LK, FF and TN compute dense deformations and our hierarchy outputs 256 predicted landmarks, from which 49 landmark locations are interpolated. The KLT tracker [13] requires temporal information and will be compared in the real video sequence.

For one image, the RMS error is computed between the estimated landmark locations $\hat{\mathbf{p}}$ and ground truth locations p as $RMS = \sqrt{\frac{1}{K}\sum_{i=1}^{K} \|\mathbf{p}(i) - \hat{\mathbf{p}}(i)\|^2}$. For multiple images, averaged RMS is reported.

Table 1 compares the performance. Due to repetitive patterns, previous approaches fail to estimate the landmarks correctly. SIFT matching fails completely. The prediction of ESR is restricted to be on the linear shape subspace spanned by the training samples. Thus, it is insufficient to use the template to capture the subspace of a complex deformation field. LK and FF are stuck in local maxima despite their coarse-to-fine implementations. Our approach obtains the best performance. Fig. 7 shows the progression of our algorithm. In terms of speed, our approach is second only to ESR, which uses a fast boosting framework.

Influence of multiple layers. It is interesting to see how the performance changes if we switch off the first L layers of predictors. As shown in Table 2, the first two layers have less contribution on the performance than the rest of the layers. On the other hand, the lower 6 layers indeed

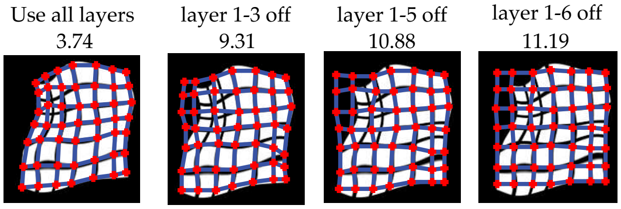


Figure 8. Performance changes if the first K layers are switched off. When more layers are switched off, the algorithm is unable to identify global deformation and is essentially the same as local template matching at each landmark.

	LK	TN	ESR	FF	SR	Ours
RMS	14.79	6.44	8.98	7.29	98.94	**5.63**
sec/frame	11	77	**0.012**	35	1.25	0.10

Table 1. Performance comparison of different approaches, including Lucas-Kanade (LK) [1], Data-driven Descent (TN) [17], Free-form registration (FF) [9], Explicit Shape Regression [3] and SIFT matching with outlier removal using RANSAC (SR) [6]. Ours is the best performer and second best in time cost per frame.

L	0	1	2	3	4	5	6
RMS	5.63	5.20	5.14	5.83	6.72	7.95	8.74

Table 2. Performance on synthetic data if the first L layer of predictors are switched off, showing the bottom layers play a critical role for performance.

help the performance. Fig. 8 demonstrates how prediction from coarse layers (large patch) help the lower layer (small patch) find correct correspondences in repetitive patterns, justifying the hierarchy.

8. Real Experiments

We also apply our framework to real world scenarios such as water distortion, cloth deformation and registration of medical images. In Fig. 9, contour tracking is achieved by interpolating contour points from frame correspondences, while the contour of the first frame is manually labeled. In Fig. 10, tracked mesh is shown.

The three water distortion sequences (Row 1-2 in Fig. 9, Row 1 in Fig. 10) and one cloth sequence (Row 3 in Fig. 9) are from [17]. Two cloth sequences (Row 2-3 in Fig. 10) are from [15] and [8]. The medical sequence of cardiac magnetic resonance images (4th row in Fig. 9) is from [18]. We captured the cloth sequence in the 5th row of Fig. 9.

For the sequences on the 4th row of Fig. 9 and the 1st row of Fig. 10, we use temporal information by adding training samples generated from perturbing the final estimation of the previous frame. This slows down the processing to 0.3-0.5fps, yet is still faster than previous approaches. For other sequences, our algorithm runs at around 3-4 fps.

Note that our method successfully estimates the deformations. In comparison, SIFT+RANSAC only obtains a sparse set of distinctive matches, not enough for estimating a nonrigid deformation (even if we are using Thin-Plate Spline). TN can capture detailed local deformations but not global shifts of the cloth without modeling the relationship between local patches. KLT trackers lose the target quickly and localize contour inaccurately.

We also quantitatively measure the landmark localization error using the densely labeled dataset provided in [17], which contains 30 labeled frames, each with 232 landmarks. In terms of RMS, LK gives 5.20, FF gives 3.93, TN gives 2.51 while our approach gives 3.29. Our framework is only second to TN, which is much slower.

We have tested our algorithm on existing datasets of deformable objects proposed by [10, 11]. Although no groundtruth is available, our performance is close to their published results (e.g. 4.10 mean pixel distance difference in cushion video [10] and 4.43 in bed-sheet video [11]). All video sequences are 404-by-504.

Acknowledgement. This research was supported in parts by ONR grant N00014-11-1-0295, a Microsoft Research PhD fellowship and a University Transportation Center T-SET grant.

References

[1] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *IJCV*, 2004. 2, 6, 7

[2] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *PAMI*, 1989. 2, 6

[3] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, 2012. 2, 6, 7

[4] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2

[5] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 2

[6] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2, 6, 7, 8

[7] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 2004. 2

[8] M. Moll and L. V. Gool. Optimal templates for non-rigid surface reconstruction. In *ECCV*, 2012. 7

[9] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, and D. Hawkes. Nonrigid registration using free-form deformations: application to breast MR images. *Medical Imaging*, 1999. 2, 6, 7

[10] M. Salzmann, R. Hartley, and P. Fua. Convex optimization for deformable surface 3-d tracking. In *ICCV*, 2007. 7

[11] M. Salzmann, F. Moreno-Noguer, V. Lepetit, and P. Fua. Closed-form solution to non-rigid 3d surface registration. In *ECCV*, 2008. 2, 7

[12] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *CVPR*, 2005. 2

[13] J. Shi and C. Tomasi. Good features to track. In *CVPR*, 1994. 2, 6, 8

[14] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011. 2

[15] J. Taylor, A. Jepson, and K. Kutulakos. Non-Rigid Structure from Locally-Rigid Motion. In *CVPR*, 2010. 7

[16] Y. Tian and S. Narasimhan. Detailed derivation of theory of hierarchical data-driven descent. *CMU RI Technical Report*, 2013. 4, 5

[17] Y. Tian and S. G. Narasimhan. Globally optimal estimation of nonrigid image distortion. *IJCV*, 2012. 1, 2, 3, 4, 5, 6, 7, 8

[18] S. Zhang, Y. Zhan, Y. Zhou, M. Uzunbas, and D. Metaxas. Shape prior modeling using sparse representation and online dictionary learning. In *Medical Image Computing and Computer-Assisted Intervention*, volume 7512 of *Lecture Notes in Computer Science*, pages 435–442. Springer Berlin Heidelberg, 2012. 7

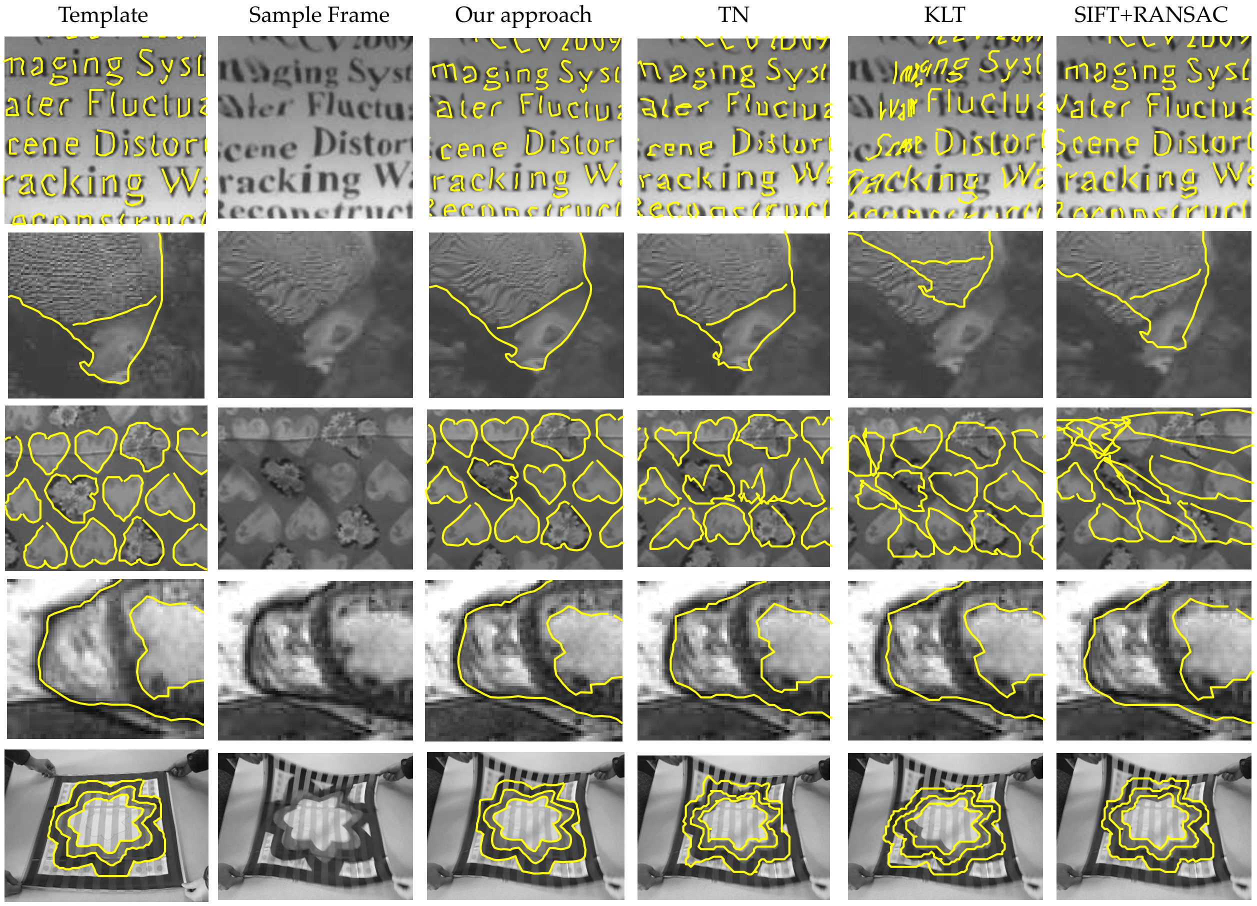


Figure 9. Example contour localization results given by our approach, TN [17], KLT [13], and SIFT matching with RANSAC [6]. Each row is a video sequence, two from underwater imaging, two from cloth deformation and the final one is from medical imaging. For each dataset, one sample frame is shown. The contours are drawn manually for the template image (1st column), and are transferred to every video frame after the correspondence was found. Our approach is stable and better than other approaches. (Best viewed in color)

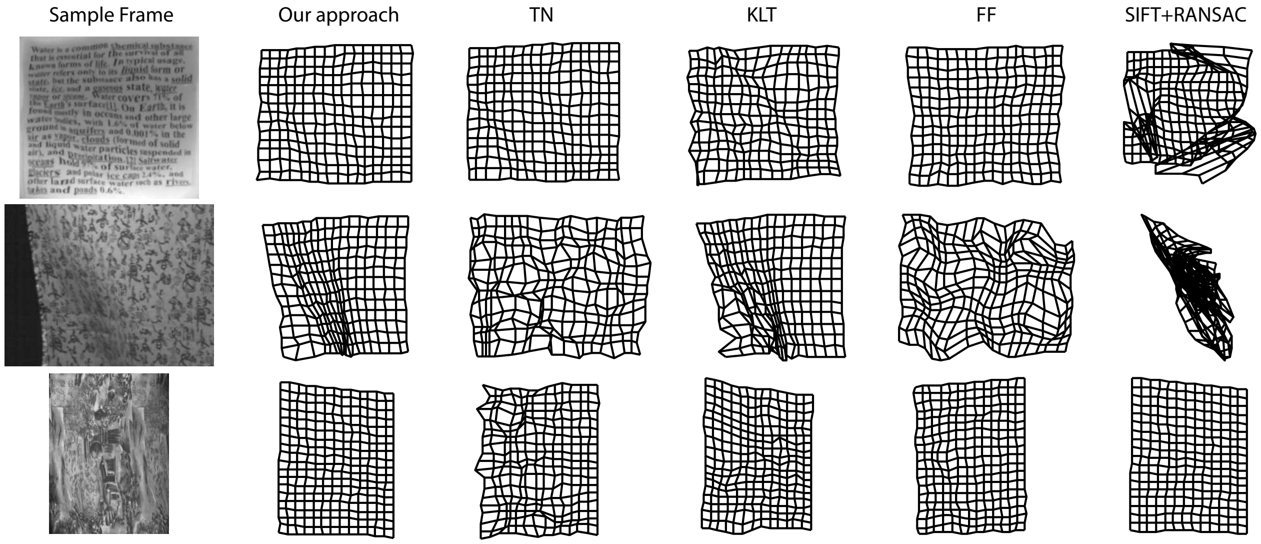


Figure 10. Example dense correspondence results given by our approach, TN, KLT, FF and SIFT matching with RANSAC. Each row is a video, two from cloth deformation and one from underwater imaging. The mesh is a regular grid on the template.