

Detailed Derivation of Theory of Hierarchical Data-driven Descent

Yuandong Tian and Srinivasa G. Narasimhan
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
{yuandong, srinivas}@cs.cmu.edu

CMU RI Technical Report

Contents

1	Image Formation Model	2
1.1	Parameterization of Deformation Field $W(\mathbf{x}; \mathbf{p})$	2
1.2	The bases function $B(x)$	2
1.3	Image Patch	2
2	Main Theorem	3
2.1	Pull-back conditions	3
2.2	Relaxed Lipchitz Conditions	4
2.3	Guaranteed Nearest Neighbor	4
2.4	Number of Samples Needed	6
3	Sampling within a Hypercube	7
3.1	Covering the Entire Hypercube	7
3.2	Covering a Subspace within Hypercube	8
4	Finding optimal curve $\gamma = \gamma(\alpha)$	10
5	More Experiments	11

1 Image Formation Model

We repeat our model here from the paper for readability. The template image T and the distorted image I_p is related by the followin equality:

$$I_p(W(\mathbf{x}; \mathbf{p})) = T(\mathbf{x}) \quad (1)$$

where, $W(\mathbf{x}; \mathbf{p})$ is the deformation field that maps the 2D location of pixel \mathbf{x} on the template with the 2D location of pixel $W(\mathbf{x}; \mathbf{p})$ on the distorted image I_p .

1.1 Parameterization of Deformation Field $W(\mathbf{x}; \mathbf{p})$

$W(\mathbf{x}; \mathbf{p})$ is parameterized by the displacements of K landmarks. Each landmark i has a rest location l_i and displacement $\mathbf{p}(i)$. Both of them are 2-dimensional column vectors. For any \mathbf{x} , its deformation $W(\mathbf{x}; \mathbf{p})$ is a weighted combination of the displacements of K landmarks:

$$W(\mathbf{x}; \mathbf{p}) = \mathbf{x} + \sum_{i=1}^K b_i(\mathbf{x}) \mathbf{p}(i) \quad (2)$$

whose $b_i(\mathbf{x})$ is the weight from landmark i to location x . Naturally we have $\sum_i b_i(\mathbf{x}) = 1$ (all weights at any location sums to 1), $b_i(l_i) = 1$ and $b_i(l_j) = 0$ for $j \neq i$. We can also write Eqn. 2 as the following matrix form:

$$W(\mathbf{x}; \mathbf{p}) = \mathbf{x} + B(\mathbf{x}) \mathbf{p} \quad (3)$$

where $B(\mathbf{x}) = [b_1(\mathbf{x}), b_2(\mathbf{x}), \dots, b_K(\mathbf{x})]$ is a K -dimensional column vector and $\mathbf{p} = [\mathbf{p}(1), \mathbf{p}(2), \dots, \mathbf{p}(K)]^\top$ is a K -by-2 matrix. Each row of \mathbf{p} is the displacement $\mathbf{p}(i)^\top$ of landmark i .

1.2 The bases function $B(x)$

Given any pixel location \mathbf{x} , the weighting function $b_i(\mathbf{x})$ satisfies $0 \leq b_i(\mathbf{x}) \leq 1$ and $\sum_i b_i(\mathbf{x}) = 1$. For landmark i , $b_i(l_i) = 1$ and $b_i(l_j) = 0$ for $j \neq i$.

We assume that $B(\mathbf{x}) = [b_1(\mathbf{x}), b_2(\mathbf{x}), \dots, b_K(\mathbf{x})]$ is smoothly changing:

Assumption 1 *There exists c_B so that:*

$$\|(B(\mathbf{x}) - B(\mathbf{y})) \mathbf{p}\|_\infty \leq c_B \|\mathbf{x} - \mathbf{y}\|_\infty \|\mathbf{p}\|_\infty \quad (4)$$

Intuitively, Eqn. 4 measures how smooth the bases change over space.

Lemma 1 (Unity bound) *For any \mathbf{x} and any \mathbf{p} , we have $\|B(\mathbf{x}) \mathbf{p}\|_\infty \leq \|\mathbf{p}\|_\infty$.*

Proof

$$\|B(\mathbf{x}) \mathbf{p}\|_\infty = \max\left\{\sum_i b_i(\mathbf{x}) \mathbf{p}^x(i), \sum_i b_i(\mathbf{x}) \mathbf{p}^y(i)\right\} \quad (5)$$

$$\leq \max\left\{\max_i \mathbf{p}^x(i) \sum_i b_i(\mathbf{x}), \max_i \mathbf{p}^y(i) \sum_i b_i(\mathbf{x})\right\} = \|\mathbf{p}\|_\infty \quad (6)$$

using the fact that $\sum_i b_i(\mathbf{x}) = 1$ for any \mathbf{x} . ■

1.3 Image Patch

We consider a square $R = R(\mathbf{x}, r) = \{\mathbf{y} : \|\mathbf{x} - \mathbf{y}\|_\infty \leq r\}$ centered at x with side $2r$. Given an image I treated as a long vector of pixels, the image content (patch) $I(R)$ is a vector obtained by selecting the components of I that is spatially contained in the square R . $S = S(\mathbf{x}, r)$ is the subset of landmarks that most influences the image content at $I(R)$. The parameters on the subset are denoted as $\mathbf{p}(S)$. Fig. 1 shows the relationship.

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 $Area(R)$, or to the square of the patch scale (r^2), which gives $2|S| \propto r^2$.

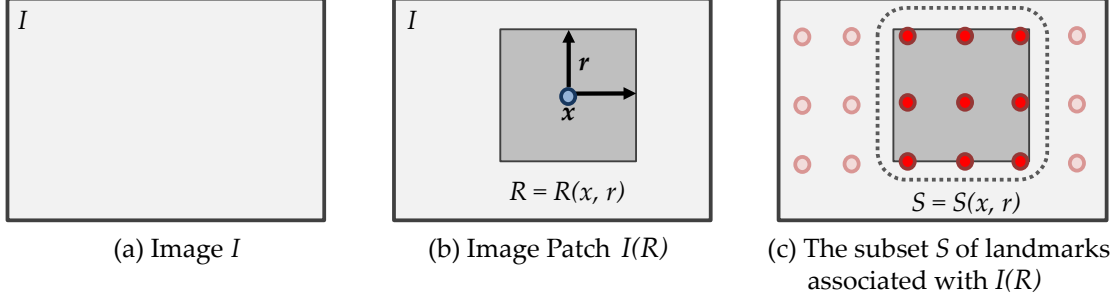


Figure 1: Notations. (a) The image I , (b) The image patch $I(R)$ centered at \mathbf{x} with side $2r$. (c) The subset S of landmarks that most influences the patch content $I(R)$.

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 . For example, if the entire image is under affine transform which has 6 degrees of freedom, then each patch $I(R)$ of that image, regardless of its scale and number of landmarks, will also under affine transform. Therefore the degrees of freedom in $I(R)$ will never exceed 6.

Given the two observations, we can assume:

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

2 Main Theorem

2.1 Pull-back conditions

Like [1], the pull-back operation takes (1) a distorted image $I_{\mathbf{p}}$ with unknown parameter \mathbf{p} and (2) parameters \mathbf{q} , and outputs an less distorted image $H(I_{\mathbf{p}}, \mathbf{q})$:

$$H(I_{\mathbf{p}}, \mathbf{q})(\mathbf{x}) \equiv I_{\mathbf{p}}(W(\mathbf{x}; \mathbf{q})) \quad (7)$$

Ideally, $H(I_{\mathbf{p}}, \mathbf{q})$ is used to simulate the appearance of $I_{\mathbf{p}-\mathbf{q}}$ without knowing the true parameters \mathbf{p} . This is indeed the case for $\mathbf{p} = \mathbf{q}$, since from Eqn. 1 we get $H(I_{\mathbf{p}}, \mathbf{p}) = T$. In general, it is not the case for $\mathbf{p} \neq \mathbf{q}$. However, the difference is bounded [1]:

$$\|H(I_{\mathbf{p}}, \mathbf{q}) - I_{\mathbf{p}-\mathbf{q}}\| \leq C_5 \|\mathbf{p} - \mathbf{q}\| \quad (8)$$

for some constant C_5 characterizing the amount of pull-back error. Similarly, we can also prove the patch version:

Theorem 2 *For patch (\mathbf{x}, r) , if $\|\mathbf{p} - \mathbf{q}\|_{\infty} \leq r$, then*

$$\|H(I_{\mathbf{p}}, \mathbf{q})(R) - I_{\mathbf{p}-\mathbf{q}}(R)\| \leq \eta(\mathbf{x}, r)r \quad (9)$$

where $\eta(\mathbf{x}, r) = c_B c_q c_G \text{Area}_j$. Note $c_G = \max_{\mathbf{y} \in R} |\nabla I_{\mathbf{p}}(\mathbf{y})|_1$, $c_q = \frac{r_1}{1-\bar{\gamma}}$ and c_B is defined in Eqn. 4.

Proof For any $\mathbf{y} \in R = R(\mathbf{x}, r)$, by definitions of Eqn. 7 and Eqn. 1, we have:

$$H(I_{\mathbf{p}}, \mathbf{q})(\mathbf{y}) = I_{\mathbf{p}}(W(\mathbf{y}; \mathbf{q})) \quad (10)$$

$$I_{\mathbf{p}-\mathbf{q}}(\mathbf{y}) = T(W^{-1}(\mathbf{y}; \mathbf{p} - \mathbf{q})) = I_{\mathbf{p}}(W(W^{-1}(\mathbf{y}; \mathbf{p} - \mathbf{q}), \mathbf{p})) \quad (11)$$

Now we need to check the pixel distance between $\mathbf{u} = W(\mathbf{y}; \mathbf{q})$ and $\mathbf{v} = W(W^{-1}(\mathbf{y}; \mathbf{p} - \mathbf{q}), \mathbf{p})$. Note both are pixel locations on distorted image $I_{\mathbf{p}}$. If we can bound $\|\mathbf{u} - \mathbf{v}\|_{\infty}$, then from $I_{\mathbf{p}}$'s appearance, we can obtain the bound for $|H(I_{\mathbf{p}}, \mathbf{q})(\mathbf{y}) - I_{\mathbf{p}-\mathbf{q}}(\mathbf{y})|$.

Denote $\mathbf{z} = W^{-1}(\mathbf{y}; \mathbf{p} - \mathbf{q})$ which is a pixel location on the template. By definition we have:

$$\mathbf{y} = \mathbf{z} + B(\mathbf{z})(\mathbf{p} - \mathbf{q}) \quad (12)$$

then we have $\|\mathbf{y} - \mathbf{z}\|_\infty = \|B(\mathbf{z})(\mathbf{p} - \mathbf{q})\|_\infty \leq \|\mathbf{p} - \mathbf{q}\|_\infty \leq r$ by Lemma 1. On the other hand, we have:

$$\mathbf{u} - \mathbf{v} = W(\mathbf{y}, \mathbf{q}) - W(\mathbf{z}, \mathbf{p}) \quad (13)$$

$$= \mathbf{y} + B(\mathbf{y})\mathbf{q} - \mathbf{z} - B(\mathbf{z})\mathbf{p} \quad (14)$$

$$= B(\mathbf{z})(\mathbf{p} - \mathbf{q}) - B(\mathbf{z})\mathbf{p} + B(\mathbf{y})\mathbf{q} \quad (15)$$

$$= (B(\mathbf{y}) - B(\mathbf{z}))\mathbf{q} \quad (16)$$

Thus, from Eqn. 4 we have:

$$\|\mathbf{u} - \mathbf{v}\|_\infty \leq c_B \|\mathbf{y} - \mathbf{z}\|_\infty \|\mathbf{q}\|_\infty \leq (c_B \|\mathbf{q}\|_\infty) r \quad (17)$$

In the algorithm, $\mathbf{q} = \hat{\mathbf{p}}^t$ is the summation of estimations from all layers 1 to $t - 1$. Therefore:

$$\|\mathbf{q}\|_\infty = \|\hat{\mathbf{p}}^t\|_\infty = \left\| \sum_{j=1}^{t-1} \tilde{\mathbf{p}}^j \right\|_\infty \leq \sum_{j=1}^{t-1} \|\tilde{\mathbf{p}}^j\|_\infty \leq \frac{r_1}{1 - \bar{\gamma}} \quad (18)$$

and is thus bounded. Thus we have:

$$\begin{aligned} |H(I_{\mathbf{p}}, \mathbf{q})(\mathbf{y}) - I_{\mathbf{p}-\mathbf{q}}(\mathbf{y})| &= |I_{\mathbf{p}}(W(\mathbf{y}; \mathbf{q})) - I_{\mathbf{p}}(W(W^{-1}(\mathbf{y}; \mathbf{p} - \mathbf{q}), \mathbf{p}))| = |I_{\mathbf{p}}(\mathbf{u}) - I_{\mathbf{p}}(\mathbf{v})| \\ &\leq |\nabla I_{\mathbf{p}}(\xi)|_1 \|\mathbf{u} - \mathbf{v}\|_\infty \end{aligned} \quad (19) \quad (20)$$

where $\xi \in \text{Line} - \text{Seg}(\mathbf{u}, \mathbf{v})$. Collecting Eqn. 20 over the entire region R gives the bound. When the algorithm runs, on the distorted image $I_{\mathbf{p}}$, the rectangle R moves from the initial location (when $\mathbf{q} = 0$) to the final destination $\mathbf{q} = \mathbf{p}$. ■

Practically the pull-back error $\eta(\mathbf{x}, r)$ is very small and can be neglected.

2.2 Relaxed Lipchitz Conditions

We put a generalized definition of relaxed Lipchitz Conditions here. The definition of relaxed Lipchitz conditions in our main paper is a special case for $\eta(\mathbf{x}, r) = 0$.

Assumption 3 (Relaxed Lipchitz Condition with pull-back error $\eta(\mathbf{x}, r) > 0$) *There exists $0 < \alpha(\mathbf{x}, r) \leq \gamma(\mathbf{x}, r) < 1$, $A(\mathbf{x}, r) > 0$ and $\Gamma(\mathbf{x}, r) > A(\mathbf{x}, r) + 2\eta(\mathbf{x}, r)$ so that for any \mathbf{p}_1 and \mathbf{p}_2 with $\|\mathbf{p}_1\|_\infty \leq r$, $\|\mathbf{p}_2\|_\infty \leq r$:*

$$\Delta \mathbf{p} \leq \alpha r \implies \Delta I \leq A r \quad (21)$$

$$\Delta \mathbf{p} \geq \gamma r \implies \Delta I \geq \Gamma r \quad (22)$$

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 $\|\mathbf{x}\|_\infty \equiv \max_i |x_i|$. The error $\eta(\mathbf{x}, r)$ is from the property of pull-back operation (See Theorem 2).

2.3 Guaranteed Nearest Neighbor

Theorem 3 (Guaranteed Nearest Neighbor for Patch j) *For any image patch (\mathbf{x}, r) , we have subset $S = S(\mathbf{x}, r)$ and image region $R = R(\mathbf{x}, r)$. Suppose we have a distorted image I so that $\|I(R) - I_{\mathbf{p}}(R)\| \leq \eta r$ with $\|\mathbf{p}\|_\infty \leq r$, then with*

$$\min \left(c_{ss} \left\lceil \frac{1}{\alpha} \right\rceil^d, \left\lceil \frac{1}{\alpha} \right\rceil^{2|S|} \right) \quad (23)$$

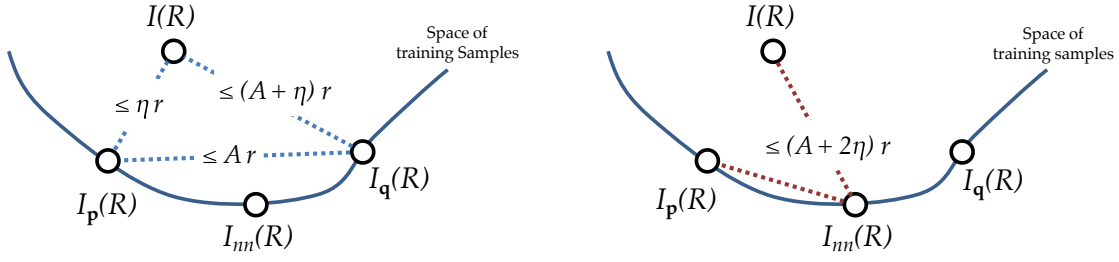


Figure 2: Illustration for proof of Guaranteed Nearest Neighbor.

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

$$\|\hat{\mathbf{p}}(S) - \mathbf{p}(S)\| \leq \gamma r \quad (24)$$

using Nearest Neighbor in the region R with image metric. Here d is the effective degrees of freedom while $2|S|$ is the apparent degrees of freedom.

Proof Since $\|\mathbf{p}\|_\infty \leq r$, by definition we have $\|\mathbf{p}(S)\|_\infty \leq r$ and similarly $\|\mathbf{q}(S)\|_\infty \leq r$. Then using Assumption 2 and applying Thm. 7 and Thm. 9, if the number of samples needed follows 23, then there exists a data sample q so that its slicing $\mathbf{q}(S)$ satisfies:

$$\|\mathbf{p}(S) - \mathbf{q}(S)\|_\infty \leq \alpha r \quad (25)$$

For $k \notin S$, the value of $\mathbf{q}(k)$ is not important as long as $\|\mathbf{q}\|_\infty \leq r$. This is because by assumption, the relaxed Lipschitz conditions still holds no matter how $\mathbf{q}(S)$ is extended to the entire landmark set.

Fig. 2 shows the relationship for different quantities involved in the proof. Consider the patch $I_p(R)$, using Eqn. 21 and we have:

$$\|I_p(R) - I_q(R)\| \leq Ar \quad (26)$$

Thus we have for the input image I :

$$\|I(R) - I_q(R)\| \leq \|I(R) - I_p(R)\| + \|I_p(R) - I_q(R)\| \leq (A + \eta)r \quad (27)$$

On the other hand, since $I_{nn}(R)$ is the Nearest Neighbor image to I , its distance to I can only be smaller:

$$\|I(R) - I_{nn}(R)\| \leq \|I(R) - I_q(R)\| \leq (A + \eta)r \quad (28)$$

Thus we have:

$$\|I_p(R) - I_{nn}(R)\| \leq \|I_p(R) - I(R)\| + \|I(R) - I_{nn}(R)\| \leq (A + 2\eta)r \quad (29)$$

Now we want to prove $\|\mathbf{p}(S) - q_{nn}(S)\| \leq \gamma r$. If not, then from Eqn. 29 we have:

$$\|I_p(R) - I_{nn}(R)\| \geq \Gamma r > (A + 2\eta)r \quad (30)$$

which from Eqn. 22 is a contradiction. Thus we have

$$\|\mathbf{p}(S) - q_{nn}(S)\|_\infty \leq \gamma r \quad (31)$$

Thus, just setting the prediction $\hat{\mathbf{p}}(S) = q_{nn}(S)$ suffices. ■

Theorem 4 (Verification of Aggregation Step.) Suppose we have estimations $\hat{\mathbf{p}}(S_j)$ for overlapping S_j of the same layer covering the same landmark i (i.e., $i \in S_j$) so that the following condition holds:

$$\|\hat{\mathbf{p}}(S_j) - p(S_j)\|_\infty \leq r \quad \forall j \quad (32)$$

Then the joint prediction

$$\tilde{p}(i) = \text{mean}_{\{j: i \in S_j\}} \hat{\mathbf{p}}_{j \rightarrow i}(S_j) \quad (33)$$

satisfies $\|\hat{\mathbf{p}}(i) - p(i)\|_\infty \leq r$. As a result, $\|\hat{\mathbf{p}} - p\|_\infty \leq r$.

Proof By the property of $\|\cdot\|_\infty$, we have for landmark i :

$$\|\hat{\mathbf{p}}_{j \rightarrow i}(S_j) - p(i)\|_\infty \leq r \quad (34)$$

Then we have

$$\|\tilde{\mathbf{p}}(i) - \mathbf{p}(i)\| = \left\| \frac{1}{\#\{j : i \in S_j\}} \sum_{j:i \in S_j} \hat{\mathbf{p}}_{j \rightarrow i}(S_j) - p(i) \right\| \leq r \quad (35)$$

■

2.4 Number of Samples Needed

Theorem 5 (The Number of Samples Needed) *The total number N of samples needed is bounded by:*

$$N \leq C_3 C_1^d + C_2 \log_{1/\bar{\gamma}} 1/\epsilon \quad (36)$$

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

Proof We divide our analysis into two cases: $d = 2K$ and $d < 2K$, where K is the number of landmarks. $d > 2K$ is not possible. We index patch (\mathbf{x}, r) with subscript j , i.e., for j -th patch, its Lipschitz constants are $\alpha_j, \gamma_j, A_j, \Gamma_j$, etc. Besides, denote $[t]$ as the subset of all patches that belong to the same layer t .

Case 1: $d = 2K$

First let us consider the case that the intrinsic dimensionality of deformation field d is just $2K$. Then the root dimensionality $d_1 = 2K$ (twice the number of landmarks). By Assumption 2, the dimensionality d_t for layer t is:

$$d_t = \beta r_t^2 = \frac{d_1}{r_1^2} r_t^2 = \bar{\gamma}^{2t-2} d_1 \quad (37)$$

Any patch $j \in [t]$ has the same degrees of freedom since by Assumption 2, d_j only depends on r_j , which is constant over layer t .

For any patch $j \in [t]$, we use at most N_j training samples:

$$N_j \leq \left(\frac{1}{\alpha_j} \right)^{d_t} \quad (38)$$

to ensure the contracting factor is indeed at least $\gamma_j \leq \bar{\gamma}$. Note for patch j , we only need the content within the region R_{j0} as the training samples. Therefore, training samples of different patches in this layer can be stitched together, yielding samples that cover the entire image. For this reason, the number N_t of training samples required for the layer t is:

$$N_t \leq \arg \max_{j \in [t]} N_j \leq C_1^{d_t} = C_1^{\bar{\gamma}^{2t-2} d_1} \quad (39)$$

for $C_1 = 1/\min_j \alpha_j$. Denote $n_t = C_1^{\bar{\gamma}^{2t-2} d_1}$. Then we have:

$$N \leq \sum_{t=1}^T N_t \leq \sum_{t=1}^T n_t \quad (40)$$

To bound this, just cut the summation into half. Given $l > 1$, set T_0 so that

$$\frac{n_{T_0}}{n_{T_0+1}} = n_{T_0}^{1-\bar{\gamma}^2} \geq l, \quad \frac{n_{T_0+1}}{n_{T_0+2}} = n_{T_0+1}^{1-\bar{\gamma}^2} \leq l \quad (41)$$

Thus we have

$$\sum_{t=1}^T n_t = \sum_{t=1}^{T_0} n_t + \sum_{t=T_0+1}^T n_t \quad (42)$$

The first summation is bounded by a geometric series. Thus we have

$$\sum_{t=1}^{T_0} n_t \leq C_1^{d_1} \sum_{t=1}^{T_0} \left(\frac{1}{l}\right)^{t-1} \leq \frac{C_1^{d_1}}{1-1/l} = \frac{l}{l-1} C_1^{d_1} \quad (43)$$

On the other hand, each item of the second summation is less than $l^{1/(1-\bar{\gamma}^2)}$. Thus we have:

$$\sum_{t=T_0+1}^T n_t \leq l^{1/(1-\bar{\gamma}^2)} T \quad (44)$$

Combining the two, we then have:

$$N \leq \frac{l}{l-1} C_1^{d_1} + l^{\frac{1}{1-\bar{\gamma}^2}} T \quad (45)$$

for $T = \lceil \log_{1/\bar{\gamma}} 1/\epsilon \rceil$. Note this bound holds for any l , e.g. 2. In this case, we have

$$N \leq 2C_1^{d_1} + C_2 T \quad (46)$$

for $C_2 = 2^{\frac{1}{1-\bar{\gamma}^2}}$.

Case 2: $d < 2K$

In this case, setting $d_1 = 2K$, finding T_1 so that $d_{T_1} \geq d$ but $d_{T_1+1} < d$ in Eqn. 37, yielding:

$$T_1 = \left\lceil \frac{1}{2} \log_{1/\bar{\gamma}} 2K/d \right\rceil + 1 \quad (47)$$

Then, by Assumption 2, from layer 1 to layer T_1 , their dimensionality is at most d . For any layer between 1 and T_1 , N_t is bounded by a constant number:

$$N_t \leq c_{ss} C_1^d \quad (48)$$

The analysis of the layers from T_1 to T follow case 1, except that we have d as the starting dimension rather than $2K$. Thus, from Eqn. 46, the total number of samples needed is:

$$N \leq (T_1 c_{ss} + 2) C_1^d + C_2 T \quad (49)$$

■

3 Sampling within a Hypercube

Theorem 3 is based on a design of sampling strategy so that for every location \mathbf{p} in the hypercube $[-r, r]^D$, there exists at least one sample sufficiently close to it. Furthermore, we want to minimize the number of samples needed for this design. Mathematically, we want to find the smallest *cover* of $[-r, r]^D$.

In the following, we provide one necessary and two sufficient conditions. The first is for the general case (covering $[-r, r]^D$ entirely), while the second specifies the number of samples needed if \mathbf{p} is known to be on a low-dimensional subspace, in which we could have better bounds.

3.1 Covering the Entire Hypercube

Theorem 6 (Sampling Theorem, Necessary Conditions) *To cover $[-r, r]^D$ with smaller hypercubes of side $2\alpha r$ ($\alpha < 1$), at least $\lfloor 1/\alpha^D \rfloor$ hypercubes are needed.*

Proof The volume of $[-r, r]^D$ is $\text{Vol}(r) = (2r)^D$, while the volume of each hypercube of side $2\alpha r$ is $\text{Vol}(2\alpha r) = (2r)^D \alpha^D$. A necessary condition of covering is the total volume of small hypercube has to be at least larger than $\text{Vol}(r)$:

$$N \text{Vol}(2\alpha r) \geq \text{Vol}(r) \quad (50)$$

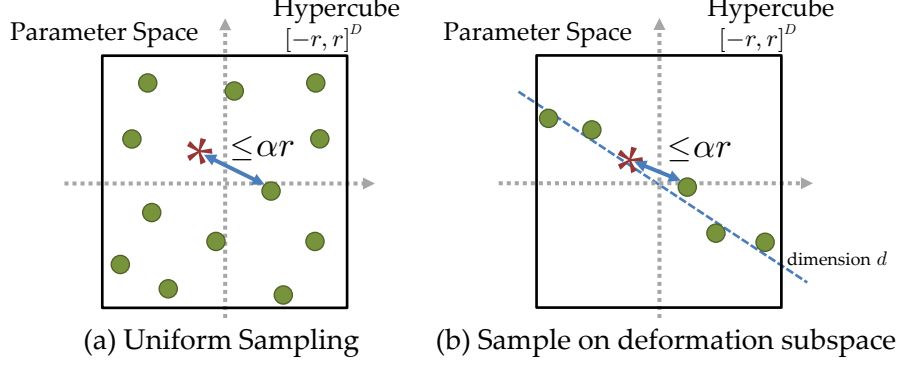


Figure 3: Sampling strategies for Thm. 7 and Thm. 9. **(a)** Uniform sampling within a hypercube $[-r, r]^D$ so that for any $\mathbf{p} \in [-r, r]^D$, there exists at least one training sample that is αr close to \mathbf{p} . **(b)** If we know that in addition to the constraint $\|\mathbf{p}\|_\infty \leq r$, \mathbf{p} always lies on a subspace of dimension $d < D$, then just assigning samples near the subspace within the hypercube suffices.

which gives:

$$N \geq \frac{\text{Vol}(r)}{\text{Vol}(2\alpha r)} = \frac{1}{\alpha^D} \geq \left\lceil \frac{1}{\alpha^D} \right\rceil \quad (51)$$

■

Theorem 7 (Sampling Theorem, Sufficient Conditions) *With $\lceil 1/\alpha \rceil^D$ number of samples ($\alpha < 1$), for any \mathbf{p} contained in the hypercube $[-r, r]^D$, there exists at least one sample $\hat{\mathbf{p}}$ so that $\|\hat{\mathbf{p}} - \mathbf{p}\|_\infty \leq \alpha r$.*

Proof Uniformly distribute the training samples within the hypercube does the job. In particular, denote

$$n = \left\lceil \frac{1}{\alpha} \right\rceil \quad (52)$$

Thus we have $1/n = 1/\lceil 1/\alpha \rceil \leq 1/(1/\alpha) = \alpha$. We put training sample of index (i_1, i_2, \dots, i_d) on d -dimensional coordinates:

$$\hat{\mathbf{p}}_{i_1, i_2, \dots, i_d} = r \left[-1 + \frac{2i_1 - 1}{n}, -1 + \frac{2i_2 - 1}{n}, \dots, -1 + \frac{2i_D - 1}{n} \right] \quad (53)$$

does the job. Here $1 \leq i_k \leq n$ for $k = 1 \dots D$. So each dimension we have n training samples. Along the dimension, the first sample is r/n distance away from $-r$, then the second sample is $2r/n$ distance to the first sample, until the last sample that is r/n distance away from the boundary r . Then for any $\mathbf{p} \in [-r, r]^D$, there exists i_k so that

$$\left| \mathbf{p}^{(k)} - r \left(-1 + \frac{2i_k - 1}{n} \right) \right| \leq \frac{1}{n} r \leq \alpha r \quad (54)$$

This holds for $1 \leq k \leq D$. As a result, we have

$$\|\mathbf{p} - \hat{\mathbf{p}}_{i_1, i_2, \dots, i_D}\|_\infty \leq \alpha r \quad (55)$$

and the total number of samples needed is $n^D = \lceil 1/\alpha \rceil^D$.

3.2 Covering a Subspace within Hypercube

Now we consider the case that \mathbf{p} lies on a subspace of dimension d , i.e., there exists a column-independent matrix U of size D -by- d so that $\mathbf{p} = U\mathbf{h}$ for some hidden variable \mathbf{h} . This happens if we use overcomplete local bases to represent the deformation. Since each landmark is related to two local bases, usually $D/2$ number of landmarks will give the deformation parameters \mathbf{p} with apparent dimension D .

In this case, we do not need to fill the entire hypercube $[-r, r]^D$. In fact, we expect the number of samples to be exponential with respect to only d rather than D .

Definition 8 (Noise Controlled Deformation Field) A deformation field \mathbf{p} is called noise-controlled deformation of order k and expanding factor c , if for every $\mathbf{p} \in [-r, r]^D$, there exists a k -dimensional vector ($k \geq d$) $\mathbf{v} \in [-r, r]^k$ so that $\mathbf{p} = f(\mathbf{v})$. Furthermore, for any $\mathbf{v}_1, \mathbf{v}_2 \in [-r, r]^k$, we have:

$$\|\mathbf{p}_1 - \mathbf{p}_2\|_\infty = \|f(\mathbf{v}_1) - f(\mathbf{v}_2)\|_\infty \leq c\|\mathbf{v}_1 - \mathbf{v}_2\|_\infty \quad (56)$$

for a constant $c \geq 1$.

Note that by the definition of intrinsic dimensionality d , \mathbf{v} could be only d -dimensional and still $\mathbf{p} = f(\mathbf{v})$. However, in this case, c could be pretty large. In order to make c smaller, we can have a *redundant* k -dimensional representation \mathbf{h} with $k > d$.

Many global deformation field satisfies Definition 8. Here we consider two cases, the affine deformation and the transformation that contains only translation and rotation.

Affine transformation. An affine deformation field \mathbf{p} defined on a grid has $d = 6$ and $k = 8$, no matter how many landmarks ($D/2$) there are. This is because each component of \mathbf{p} can be written as

$$\mathbf{p}(k) = [\lambda_1 x_k + \lambda_2 y_k + \lambda_3, \lambda_4 x_k + \lambda_5 y_k + \lambda_6] \quad (57)$$

for location $\mathbf{l}_k = (x_k, y_k)$. Therefore, since any landmarks \mathbf{l}_k within a rectangle can be linearly represented by the locations of four corners in a convex manner, the deformation vector $\mathbf{p}(k)$ on \mathbf{l}_k can also be linearly represented by the deformation vectors of four corners (8 DoF):

$$\mathbf{p}(k) = A_k \mathbf{v} = \sum_{j=1}^4 a_{kj} \mathbf{v}(j) \quad (58)$$

with \mathbf{v} is the concatenation of four deformation vectors from the four corners, $0 \leq a_{kj} \leq 1$ and $\sum_j a_{kj} = 1$. For any $\mathbf{p} \in [-r, r]^D$, \mathbf{v} can be found by just picking the deformation of its four corners, and thus $\|\mathbf{v}\|_\infty \leq r$. Furthermore, we have for $\mathbf{v}_1, \mathbf{v}_2 \in [-r, r]^k$:

$$\|\mathbf{p}_1 - \mathbf{p}_2\|_\infty = \|f(\mathbf{v}_1) - f(\mathbf{v}_2)\|_\infty \leq \max_k \sum_{j=1}^4 a_{kj} \|\mathbf{v}_1(j) - \mathbf{v}_2(j)\| \leq \|\mathbf{v}_1 - \mathbf{v}_2\|_\infty \quad (59)$$

Therefore, $c = 1$.

Transformation that contains only translation and rotation. Similarly, for deformation that contains pure translation and rotation ($d = 3$), we just pick displacement vectors on two points ($k = 4$), the rotation center and the corner as \mathbf{v} . Then we have:

$$\mathbf{p}(r, \theta) = \mathbf{p}_{\text{center}} + \frac{r}{r_{\text{corner}}} R(\theta) (\mathbf{p}_{\text{corner}} - \mathbf{p}_{\text{center}}) \quad (60)$$

$$= \left(I - \frac{r}{r_{\text{corner}}} R(\theta)\right) \mathbf{p}_{\text{center}} + \frac{r}{r_{\text{corner}}} R(\theta) \mathbf{p}_{\text{corner}} \quad (61)$$

where I is the identity matrix, $R(\theta)$ is the 2D rotational matrix and r_{corner} is the distance from the center location to the corner. Here we reparameterize the landmarks with polar coordinates (r, θ) . Therefore, for two different \mathbf{v}_1 and \mathbf{v}_2 , since $r \leq r_{\text{corner}}$, we have:

$$\|\mathbf{p}_1(r, \theta) - \mathbf{p}_2(r, \theta)\|_\infty \leq \left\| \left(I - \frac{r}{r_{\text{corner}}} R(\theta)\right) (\mathbf{p}_{\text{center},1} - \mathbf{p}_{\text{center},2}) \right\|_\infty \quad (62)$$

$$+ \left\| \frac{r}{r_{\text{corner}}} R(\theta) (\mathbf{p}_{\text{corner},1} - \mathbf{p}_{\text{corner},2}) \right\|_\infty \quad (63)$$

$$\leq 2\|\mathbf{p}_{\text{center},1} - \mathbf{p}_{\text{center},2}\|_\infty + \sqrt{2}\|\mathbf{p}_{\text{corner},1} - \mathbf{p}_{\text{corner},2}\|_\infty \quad (64)$$

$$\leq (2 + \sqrt{2})\|\mathbf{v}_1 - \mathbf{v}_2\|_\infty \quad (65)$$

since $|\cos(\theta)| + |\sin(\theta)| \leq \sqrt{2}$. Therefore,

$$\|\mathbf{p}_1 - \mathbf{p}_2\|_\infty = \max_{r, \theta} \|\mathbf{p}_1(r, \theta) - \mathbf{p}_2(r, \theta)\|_\infty \leq (2 + \sqrt{2})\|\mathbf{v}_1 - \mathbf{v}_2\|_\infty \quad (66)$$

So $c = 2 + \sqrt{2} \leq 3.5$.

Given this definition, we thus have the following sampling theorem for deformation parameters \mathbf{p} lying on a subspace that is noise-controlled.

Theorem 9 (Sampling Theorem, Sufficient Condition for Subspace Case) *For any noise-controlled deformation field $\mathbf{p} = f(\mathbf{v})$ with order k and expanding factor c , with $c_{ss} \lceil 1/\alpha \rceil^d$ number of training samples distributed in the hypercube $[-r, r]^D$, there exists at least one sample $\hat{\mathbf{p}}$ so that $\|\hat{\mathbf{p}} - \mathbf{p}\|_\infty \leq \alpha r$. Note $c_{ss} = \lceil c \rceil^k \lceil \frac{1}{\alpha} \rceil^{k-d}$.*

Proof We first apply Thm. 7 to the hypercube $[-r, r]^k$. Then with $\lceil \frac{c}{\alpha} \rceil^k$ samples, for any $\mathbf{v} \in [-r, r]^k$, there exists a training sample \mathbf{v}^i so that

$$\|\mathbf{v} - \mathbf{v}^i\|_\infty \leq \frac{\alpha r}{c} \quad (67)$$

We then build the training samples $\{\mathbf{p}^i\}$ by setting $\mathbf{p}^i = f(\mathbf{v}^i)$. Therefore, from the definition of noise cancelling, given any $\mathbf{p} \in [-r, r]^D$, there exists an $\mathbf{v} \in [-r, r]^k$ so that $\mathbf{p} = f(\mathbf{v})$. By the sampling procedure, there exists \mathbf{v}^i so that $\|\mathbf{v} - \mathbf{v}^i\|_\infty \leq \frac{\alpha}{c}r$, and therefore:

$$\|\mathbf{p} - \mathbf{p}^i\|_\infty \leq c\|\mathbf{v} - \mathbf{v}^i\|_\infty \leq \alpha r \quad (68)$$

setting $\hat{\mathbf{p}} = \mathbf{p}^i$ thus does the job. Finally, note that

$$\left\lceil \frac{c}{\alpha} \right\rceil^k \leq \lceil c \rceil^k \left\lceil \frac{1}{\alpha} \right\rceil^{k-d} \left\lceil \frac{1}{\alpha} \right\rceil^d \quad (69)$$

So setting $c_{ss} = \lceil c \rceil^k \lceil \frac{1}{\alpha} \rceil^{k-d}$ suffices (since $\lceil ab \rceil \leq \lceil a \rceil \lceil b \rceil$). \blacksquare

4 Finding optimal curve $\gamma = \gamma(\alpha)$

Without loss of generality, we set $r = 1$. Then, we rephrase the algorithm in Alg. 1.

Algorithm 1 Find Local Lipschitz Constants

- 1: **INPUT** Parameter distances $\{\Delta \mathbf{p}_m\}$ with $\Delta \mathbf{p}_m \leq \Delta \mathbf{p}_{m+1}$.
 - 2: **INPUT** Image distances $\{\Delta I_m\}$.
 - 3: **INPUT** Scale r and noise η .
 - 4: $\Delta I_m^+ = \max_{1 \leq l \leq m} \Delta I_l$, for $i = 1 \dots M$.
 - 5: $\Delta I_m^- = \min_{i \leq l \leq M} \Delta I_l$, for $i = 1 \dots M$.
 - 6: **for** $m = 1$ to M **do**
 - 7: Find minimal $l^* = l^*(m)$ so that $\Delta I_{l^*}^- > \Delta I_m^+ + 2\eta$.
 - 8: **if** $m \leq l^*$ **then**
 - 9: Store the 4-tuples $(\alpha, \gamma, A, \Gamma) = (\Delta \mathbf{p}_m, \Delta \mathbf{p}_{l^*}, \Delta I_m^+, \Delta I_{l^*}^-)/r$.
 - 10: **end if**
 - 11: **end for**
-

To analyze Alg. 1, we make the following definitions:

Definition 10 (Allowable set of A and Γ) *Given α , define the allowable set $\tilde{A}(\alpha)$ as:*

$$\tilde{A}(\alpha) = \{A : \forall m \Delta \mathbf{p}_m \leq \alpha \implies \Delta I_m \leq A\} \quad (70)$$

Naturally we have $\tilde{A}(\alpha') \subset \tilde{A}(\alpha)$ for $\alpha' > \alpha$. Similarly, given γ , define the allowable set $\tilde{\Gamma}(\gamma)$ as:

$$\tilde{\Gamma}(\gamma) = \{\Gamma : \forall m \Delta \mathbf{p}_m \geq \gamma \implies \Delta I_m \geq \Gamma\} \quad (71)$$

and $\tilde{\Gamma}(\gamma') \subset \tilde{\Gamma}(\gamma)$ for $\gamma' < \gamma$.

Lemma 11 (Properties of ΔI^+ and ΔI^-) *The two arrays constructed in Alg. 1 satisfy:*

$$\Delta I_m^+ = \min \tilde{A}(\Delta \mathbf{p}_m) \quad (72)$$

$$\Delta I_m^- = \max \tilde{\Gamma}(\Delta \mathbf{p}_m) \quad (73)$$

Moreover, ΔI_m^+ is ascending while ΔI_m^- is descending with respect to $1 \leq m \leq M$.

Proof (a): First we show $\Delta I_m^+ \in \tilde{A}(\Delta \mathbf{p}_m)$. Since the list $\{\Delta \mathbf{p}_m\}$ was ordered, for any $\Delta \mathbf{p}_l \leq \Delta \mathbf{p}_m$, we have $l \leq m$. By definition of ΔI_m^+ , we have $\Delta I_l \leq \Delta I_m^+$. Thus $\Delta I_m^+ \in \tilde{A}(\Delta \mathbf{p}_m)$.

(b): Then we show for any $A \in \tilde{A}(\Delta \mathbf{p}_m)$, $\Delta I_m^+ \leq A$. For any $1 \leq l \leq m$, since $\Delta \mathbf{p}_l \leq \Delta \mathbf{p}_m$, by the definition of A , we have $\Delta I_l \leq A$, and thus $\Delta I_m^+ = \max_{1 \leq l \leq m} \Delta I_l \leq A$.

Therefore, $\Delta I_m^+ = \min \tilde{A}(\Delta \mathbf{p}_m)$. Similarly we can prove $\Delta I_m^- = \max \tilde{\Gamma}(\Delta \mathbf{p}_m)$. ■

Theorem 12 For each $\alpha = \Delta \mathbf{p}_m$, Algorithm 1 without the check $\alpha \leq \gamma$ always gives the globally optimal solution to the following linear programming:

$$\min \quad \gamma \quad (74)$$

$$\text{s.t.} \quad \Delta I_m \leq A \quad \forall \Delta \mathbf{p}_m \leq \alpha \quad (\text{or } A \in \tilde{A}(\alpha)) \quad (75)$$

$$\Delta I_m \geq \Gamma \quad \forall \Delta \mathbf{p}_m \geq \gamma \quad (\text{or } \Gamma \in \tilde{\Gamma}(\gamma)) \quad (76)$$

$$A + 2\eta < \Gamma \quad (77)$$

which has at least one feasible solution ($A \rightarrow +\infty, \gamma \rightarrow -\infty, \Gamma \rightarrow -\infty$) for any α .

Proof Since there are M data points, we can discretize the values of α and γ into M possible values without changing the property of solution.

(a) First we prove every solution given by Alg. 1 (without the final check) is a feasible solution to the optimization (Eqn. 74). Indeed, for any $\alpha = \Delta \mathbf{p}_m$, according to Lemma 11, $A = \Delta I_m^+ \in \tilde{A}(\alpha)$, $\gamma = \Delta \mathbf{p}_{l^*}$, and $\Gamma = \Delta I_{l^*}^- \in \tilde{\Gamma}(\gamma)$ and thus Eqn. 75 and Eqn. 76 are satisfied. From the construction of Alg. 1, $A + 2\eta < \Gamma$. Thus, the Algorithm 1 gives a feasible solution to Eqn. 74.

(b) Then we prove Alg. 1 (without the final check) gives the optimal solution. If there exists $l' < l^*$ so that $\gamma' = \Delta \mathbf{p}_{l'} < \Delta \mathbf{p}_{l^*} = \gamma$ is part of a better solution $(\alpha, \gamma', A', \Gamma')$, then $\tilde{\Gamma}(\gamma') \subset \tilde{\Gamma}(\gamma)$. This means

$$A' + 2\eta < \Gamma' \leq \Delta I_{l'}^- = \max \tilde{\Gamma}(\gamma') \leq \max \tilde{\Gamma}(\gamma) = \Delta I_{l^*}^- \quad (78)$$

On the other hand, $A = \Delta I_m^+ = \min \tilde{A}(\alpha) \leq A' \in \tilde{A}(\alpha)$. Then, there are two cases:

- $\Delta I_m^+ + 2\eta < \Delta I_{l'}^- < \Delta I_{l^*}^-$. This is not possible since the algorithm already find the minimal l^* .
- $\Delta I_m^+ + 2\eta < \Delta I_{l'}^- = \Delta I_{l^*}^-$. Then according to the algorithm, $l' = l^*$.

which is a contradiction. ■

From Theorem 12, it is thus easy to check that the complete Algorithm 1 (with the check $\alpha \leq \gamma$) gives the optimal pair (α, γ) that satisfies the Relaxed Lipschitz Conditions (Eqn. 21 and Eqn. 22).

5 More Experiments

Fig. 4 shows the behaviors of our algorithm over different iterations. We can see with more and more stages, the estimation captures more detailed structures and becomes better.

Fig. 5 shows how the performance degrades if only the bottom K layers are used for prediction. We can see that each layer plays a different rule. Layer 3-4 seems to be critical for the synthetic data since they have captured the major mode/scale of deformation.

References

- [1] Y. Tian and S. G. Narasimhan. Globally optimal estimation of nonrigid image distortion. *IJCV*, 2012.

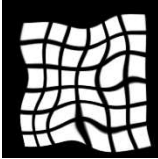
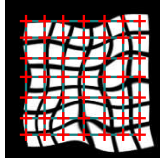
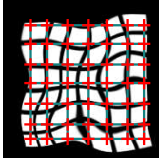
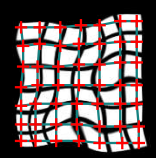
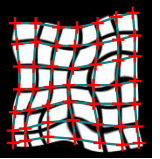
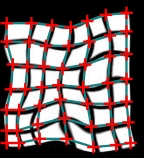
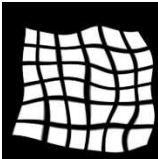
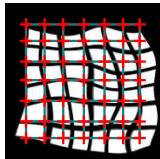
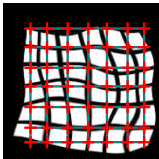
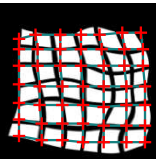
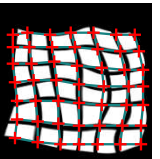
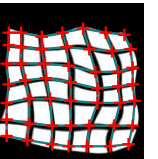
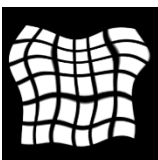
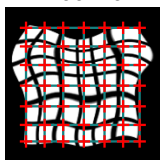
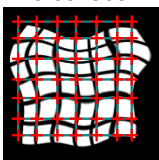
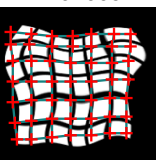
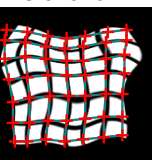
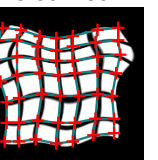

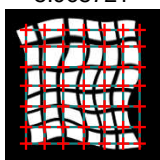
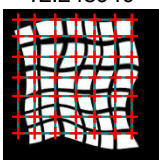
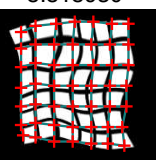
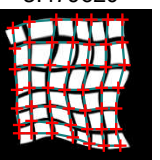
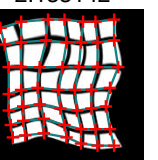
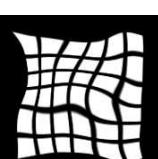
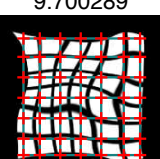
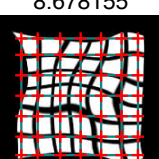
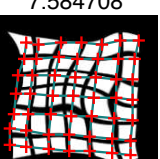
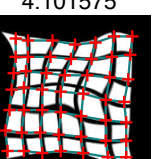
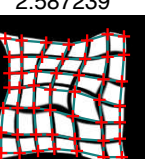

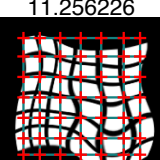
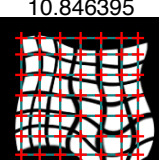
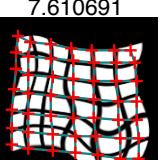

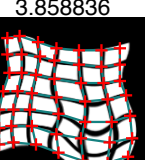

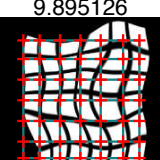
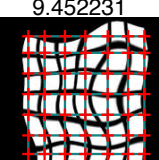
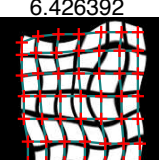
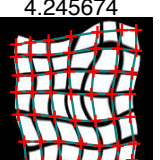
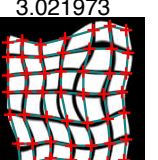
Test Image	Initialization	Iteration 1	Iteration 3	Iteration 5	Final Result
	10.934074 	9.694556 	6.870912 	4.065068 	2.555294 
	10.830096 	8.749503 	6.417697 	4.283149 	2.577389 
	12.664164 	9.052568 	7.401009 	5.070467 	3.802439 
	8.968721 	12.248940 	5.515930 	3.479620 	2.133142 
	9.700289 	8.678155 	7.584708 	4.101575 	2.587239 
	11.256226 	10.846395 	7.610691 	5.212512 	3.858836 
	9.895126 	9.452231 	6.426392 	4.245674 	3.021973 

Figure 4: Landmark Estimation at different iterations given by our approach.

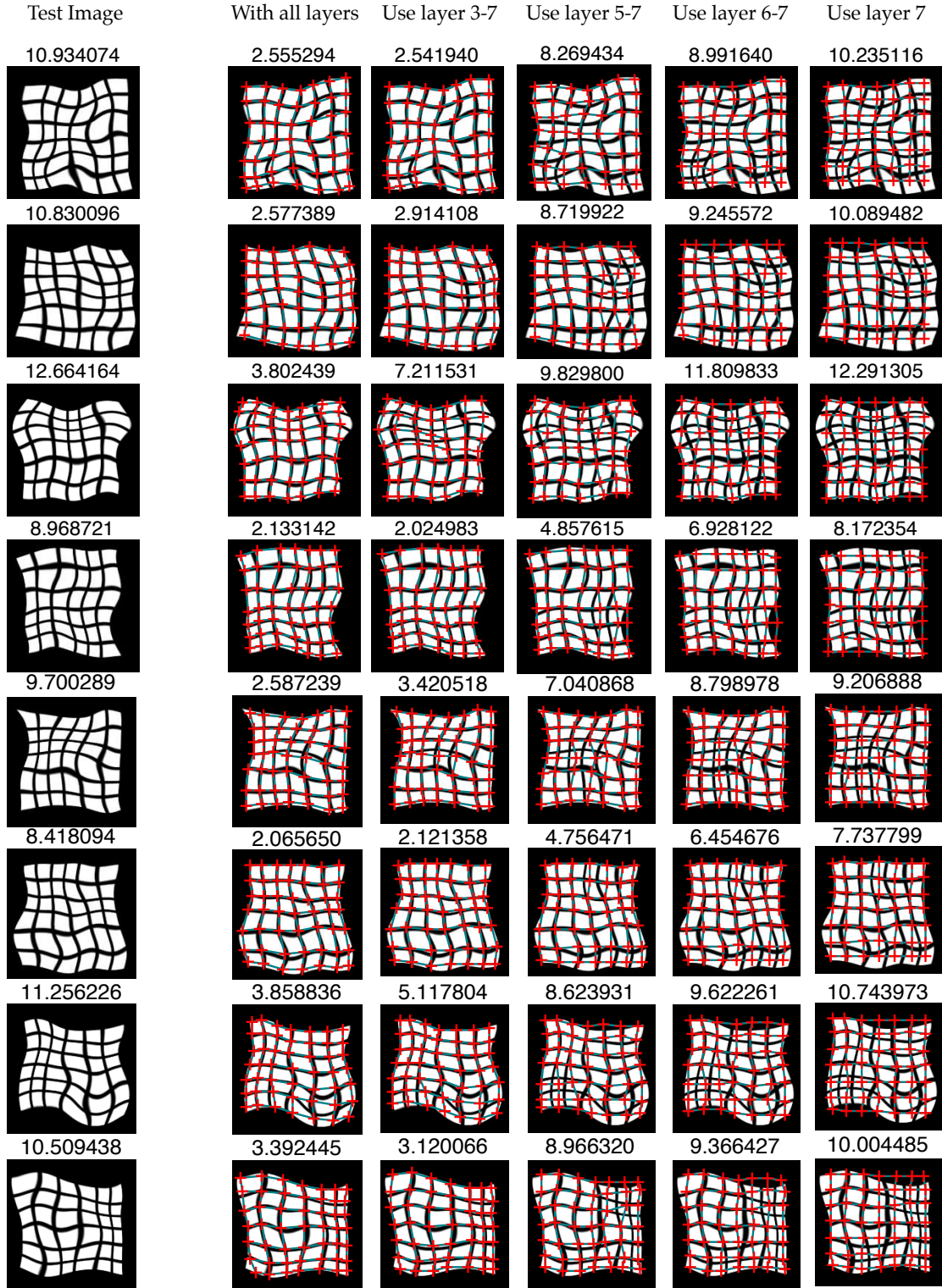


Figure 5: Landmark Estimation using only last L layers of the hierarchy. Layer 3-4 is critical for getting a good estimation of the landmarks on the synthetic data.