Supplementary Material for Paper Submission 307

Anonymous Author(s)

October 12, 2013

Contents

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:

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

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

1.1 Parameterization of Deformation Field $W(x;p)$

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

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

whose $b_i(x)$ is the weight from landmark i to location x. Naturally we have $\sum_i b_i(x) = 1$ (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:

$$W(x,p) = x + B(x)p \tag{3}$$

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

1.2 The bases function $B(x)$

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

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

Assumption 1 *There exists c_B so that:*

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

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

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

Proof

$$\|B(x)p\|_\infty = \max\{\sum_i b_i(x)p_i^x, \sum_i b_i(x)p_i^y\} \leq \max\{\max_i p_i^x \sum_i b_i(x), \max_i p_i^y \sum_i b_i(x)\} = \|p\|_\infty \tag{5}$$

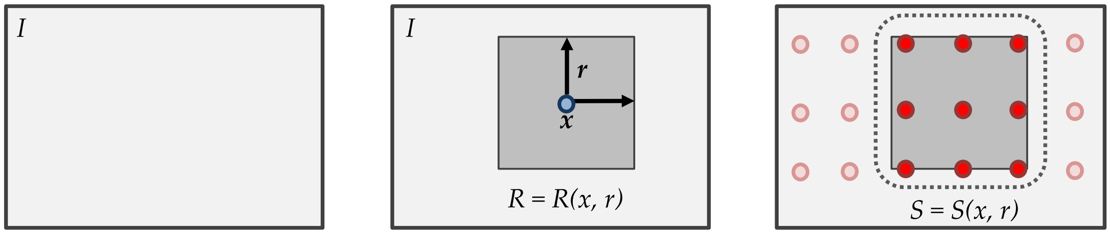
using the fact that $\sum_i b_i(x) = 1$ for any x. ∎

1.3 Image Patch

We consider a square $R = R(x,r) = \{y : \|x - y\|_\infty \leq r\}$ 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. $S = S(x,r)$ is the subset of landmarks that most influences the image content at $I(R)$. The parameters on the subset are denoted as $p(S)$. Fig. 1 shows the relationship.

Since $p(S)$ 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 $2|S| \propto r^2$.

On the other hand, if the overall effective degree of freedom is d, then no matter how large $2|S|$ is, $p(S)$ contains dependent displacements and the effective degree of freedom in R never



(a) Image I (b) Image Patch $I(R)$ (c) The subset S of landmarks associated with $I(R)$

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

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* (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_p with unknown parameter p and (2) parameters q, and outputs an less distorted image $H(I_p, q)$:

$$H(I_p, q)(x) \equiv I_p(W(x; q)) \tag{6}$$

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

$$\|H(I_p, q) - I_{p-q}\| \leq C_5 \|p - q\| \tag{7}$$

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

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

$$\|H(I_p, q)(R) - I_{p-q}(R)\| \leq \eta(x, r) r \tag{8}$$

where $\eta(x, r) = c_B c_q c_G Area_j$*. Note* $c_G = \max_{x \in R} |\nabla I_p(x)|_1$*,* $c_q = \frac{r_1}{1-\bar{\gamma}}$ *and* c_B *is defined in Eqn. 4.*

Proof For any $y \in R = R(x, r)$, by definitions of Eqn. 6 and Eqn. 1, we have:

$$\begin{aligned} H(I_p, q)(\mathbf{y}) &= I_p(W(y; q)) & (9)\\ I_{p-q}(\mathbf{y}) &= T(W^{-1}(y; p - q)) = I_p(W(W^{-1}(y; p - q), p)) & (10) \end{aligned}$$

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

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

$$y = z + B(z)(p - q) \tag{11}$$

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

$$\begin{aligned}
u - v &= W(y,q) - W(z,p) && (12)\\
&= y + B(y)q - z - B(z)p && (13)\\
&= B(z)(p-q) - B(z)p + B(y)q && (14)\\
&= (B(y) - B(z))q && (15)
\end{aligned}$$

Thus, from Eqn. 4 we have:

$$\|u-v\|_\infty \leq c_B \|y-z\|_\infty \|q\|_\infty \leq (c_B\|q\|_\infty)r \tag{16}$$

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

$$\|q\|_\infty = \|\hat{p}^t\|_\infty = \|\sum_{j=1}^{t-1} \tilde{p}^j\| \leq \sum_{j=1}^{t-1} \|\tilde{p}^j\| \leq \frac{r_1}{1-\bar{\gamma}} \tag{17}$$

and is thus bounded. Thus we have:

$$\begin{aligned}
|H(I_p, q)(y) - I_{p-q}(y)| &= |I_p(W(y;q)) - I_p(W(W^{-1}(y;p-q),p))| = |I_p(u) - I_p(v)| && (18)\\
&\leq |\nabla I_p(\xi)|_1 \|u-v\|_\infty && (19)
\end{aligned}$$

where $\xi \in Line-Seg(u,v)$. Collecting Eqn. 19 over the entire region R gives the bound. When the algorithm runs, on the distorted image I_p, the rectangle R moves from the initial location (when $q=0$) to the final destination $q=p$. ∎

Practically the pull-back error $\eta(x,r)$ 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 $\eta(x,r) = 0$.

Assumption 3 (Relaxed Lipchitz Condition with pull-back error $\eta(x,r) > 0$) *There exists $0 < \alpha(x,r) \leq \gamma(x,r) < 1$, $A(x,r) > 0$ and $\Gamma(x,r) > A(x,r) + 2\eta(x,r)$ so that for any p_1 and p_2 with $\|p_1\|_\infty \leq r$, $\|p_2\|_\infty \leq r$:*

$$\begin{aligned}
\Delta p \leq \alpha r \quad &\Longrightarrow \quad \Delta I \leq A r && (20)\\
\Delta p \geq \gamma r \quad &\Longrightarrow \quad \Delta I \geq \Gamma r && (21)
\end{aligned}$$

for $\Delta p \equiv \|p_1(S) - p_2(S)\|_\infty$ and $\Delta I \equiv \|I_{p_1}(R) - I_{p_2}(R)\|$.

Here $\|x\|_\infty \equiv \max_i |x_i|$. The error $\eta(x,r)$ 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 (x,r), we have subset $S = S(x,r)$ and image region $R = R(x,r)$. Suppose we have a distorted image I so that $\|I(R) - I_p(R)\| \leq \eta r$ with $\|p\|_\infty \leq r$, then with*

$$\min\left(c_{SS}\left\lceil\frac{1}{\alpha}\right\rceil^d, \left\lceil\frac{1}{\alpha}\right\rceil^{2|S|}\right) \tag{22}$$

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

$$\|\hat{p}(S) - p(S)\| \leq \gamma r \tag{23}$$

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.

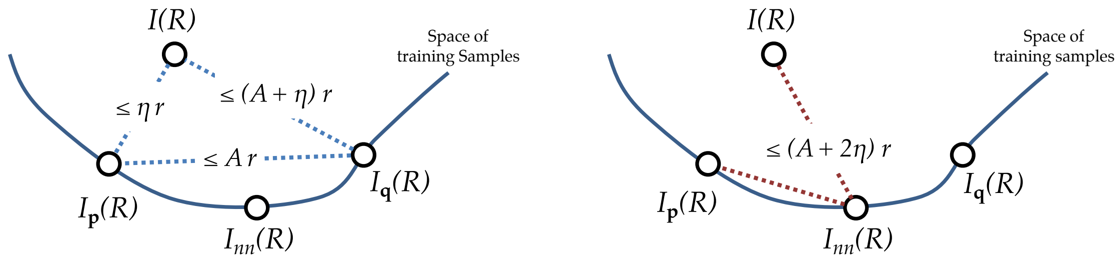


Figure 2: Illustration for proof of Guaranteed Nearest Neighbor.

Proof Since $\|p\|_\infty \le r$, by definition we have $\|p(S)\|_\infty \le r$ and similarly $\|q(S)\|_\infty \le r$. Then using Assumption 2 and applying Thm. 7 and Thm. 9, if the number of samples needed follows 22, then there exists a data sample q so that its slicing $q(S)$ satisfies:

$$\|p(S) - q(S)\|_\infty \le \alpha r \tag{24}$$

For $k \notin S$, the value of $q(k)$ is not important as long as $\|q\|_\infty \le r$. This is because by assumption, the relaxed Lipschitz conditions still holds no matter how $q(S)$ 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. 20 and we have:

$$\|I_p(R) - I_q(R)\| \le Ar \tag{25}$$

Thus we have for the input image I:

$$\|I(R) - I_q(R)\| \le \|I(R) - I_p(R)\| + \|I_p(R) - I_q(R)\| \le (A + \eta)r \tag{26}$$

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

$$\|I(R) - I_{nn}(R)\| \le \|I(R) - I_{\mathbf{q}}(R)\| \le (A + \eta)r \tag{27}$$

Thus we have:

$$\|I_p(R) - I_{nn}(R)\| \le \|I_p(R) - I(R)\| + \|I(R) - I_{nn}(R)\| \le (A + 2\eta)r \tag{28}$$

Now we want to prove $\|p(S) - q_{nn}(S)\| \le \gamma r$. If not, then from Eqn. 28 we have:

$$\|I_p(R) - I_{nn}(R)\| \ge \Gamma r > (A + 2\eta)r \tag{29}$$

which from Eqn. 21 is a contradiction. Thus we have

$$\|p(S) - q_{nn}(S)\|_\infty \le \gamma r \tag{30}$$

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

Theorem 4 (Verification of Aggregation Step.) *Supose we have estimations $\hat{p}(S_j)$ 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:*

$$\|\hat{p}(S_j) - p(S_j)\|_\infty \le r \quad \forall j \tag{31}$$

Then the joint prediction

$$\tilde{p}(i) = \mathrm{mean}_{\{j: i \in S_j\}} \hat{p}_{j \to i}(S_j) \tag{32}$$

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

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

$$\|\hat{p}_{j \to i}(S_j) - p(i)\|_\infty \le r \tag{33}$$

Then we have

$$\|\tilde{p}(i) - p(i)\| = \left\| \frac{1}{\#\{j : i \in S_j\}} \sum_{j: i \in S_j} \hat{p}_{j \to i}(S_j) - p(i) \right\| \le r \tag{34}$$

∎

2.4 Number of Samples Needed

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

$$N \le C_3 C_1^d + C_2 \log_{1/\bar{\gamma}} 1/\epsilon \tag{35}$$

where $C_1 = 1/\min \alpha(x, r)$, $C_2 = 2^{1/(1-\bar{\gamma}^2)}$ *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 (x, r) with subscript j, i.e., for j-th patch, its Lipschitz constants are α_j, γ_j, A_j, Γ_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:

$$d_t = \beta r_t^2 = \frac{d_1}{r_1^2} r_t^2 = \bar{\gamma}^{2t-2} d_1 \tag{36}$$

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:

$$N_j \le \left(\frac{1}{\alpha_j}\right)^{d_t} \tag{37}$$

to ensure the contracting factor is indeed at least $\gamma_j \le \bar{\gamma}$. 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:

$$N_t \le \arg\max_{j \in [t]} N_j \le C_1^{d_t} = C_1^{\bar{\gamma}^{2t-2} d_1} \tag{38}$$

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

$$N \le \sum_{t=1}^{T} N_t \le \sum_{t=1}^{T} n_t \tag{39}$$

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

$$\frac{n_{T_0}}{n_{T_0+1}} = n_{T_0}^{1-\bar{\gamma}^2} \ge l, \quad \frac{n_{T_0+1}}{n_{T_0+2}} = n_{T_0+1}^{1-\bar{\gamma}^2} \le l \tag{40}$$

Thus we have

$$\sum_{t=1}^{T} n_t = \sum_{t=1}^{T_0} n_t + \sum_{t=T_0+1}^{T} n_t \tag{41}$$

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

$$\sum_{t=1}^{T_0} n_t \le C_1^{d_1} \sum_{t=1}^{T_0} \left(\frac{1}{l}\right)^{t-1} \le \frac{C_1^{d_1}}{1 - 1/l} = \frac{l}{l-1} C_1^{d_1} \tag{42}$$

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

$$\sum_{t=T_0+1}^{T} n_t \le l^{1/(1-\bar{\gamma}^2)} T \tag{43}$$

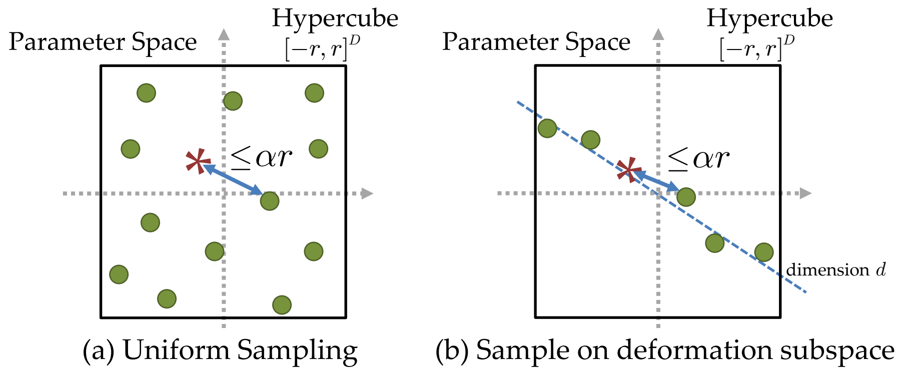


Figure 3: Sampling strategies for Thm. 7 and Thm. 9. **(a)** Uniform sampling within a hypercube $[-r, r]^D$ so that for any $\mathbf{p} \in [-r, r]^D$, there exists at least one training sample that is αr close to $\mathbf{p}$. **(b)** If we know that in addition to the constraint $\|\mathbf{p}\|_\infty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

Combining the two, we then have:

$$N \leq \frac{l}{l-1} C_1^{d_1} + l^{\frac{1}{1-\bar{\gamma}^2}} T \tag{44}$$

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

$$N \leq 2C_1^{d_1} + C_2 T \tag{45}$$

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. 36, yielding:

$$T_1 = \left\lceil \frac{1}{2} \log_{1/\bar{\gamma}} 2K/d \right\rceil + 1 \tag{46}$$

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:

$$N_t \leq c_{SS} C_1^d \tag{47}$$

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. 45, the total number of samples needed is:

$$N \leq (T_1 c_{SS} + 2) C_1^d + C_2 T \tag{48}$$

∎

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

$$N\mathrm{Vol}(2\alpha r) \geq \mathrm{Vol}(r) \tag{49}$$

which gives:

$$N \geq \frac{\mathrm{Vol}(r)}{\mathrm{Vol}(2\alpha r)} = \frac{1}{\alpha^D} \geq \left\lfloor \frac{1}{\alpha^D} \right\rfloor \tag{50}$$

∎

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

$$n = \left\lceil \frac{1}{\alpha} \right\rceil \tag{51}$$

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

$$\hat{\mathbf{p}}_{i_1,i_2,\ldots,i_d} = r\left[-1 + \frac{2i_1 - 1}{n}, -1 + \frac{2i_2 - 1}{n}, \ldots, -1 + \frac{2i_D - 1}{n}\right] \tag{52}$$

does the job. Here $1 \leq i_k \leq n$ for $k = 1 \ldots D$. 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

$$\left|\mathbf{p}(k) - r\left(-1 + \frac{2i_k - 1}{n}\right)\right| \leq \frac{1}{n} r \leq \alpha r \tag{53}$$

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

$$\|\mathbf{p} - \hat{\mathbf{p}}_{i_1,i_2,\ldots,i_D}\|_\infty \leq \alpha r \tag{54}$$

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:*

$$\|\mathbf{p}_1 - \mathbf{p}_2\|_\infty = \|f(\mathbf{v}_1) - f(\mathbf{v}_2)\|_\infty \leq c\|\mathbf{v}_1 - \mathbf{v}_2\|_\infty \tag{55}$$

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

$$\mathbf{p}(k) = [\lambda_1 x_k + \lambda_2 y_k + \lambda_3, \lambda_4 x_k + \lambda_5 y_k + \lambda_6] \tag{56}$$

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):

$$\mathbf{p}(k) = A_k \mathbf{v} = \sum_{j=1}^{4} a_{kj} \mathbf{v}(j) \tag{57}$$

with $\mathbf{v}$ is the concatenation of four deformation vectors from the four corners, $0 \le a_{kj} \le 1$ 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 $\|\mathbf{v}\|_\infty \le r$. Furthermore, we have for $\mathbf{v}_1, \mathbf{v}_2 \in [-r, r]^k$:

$$\|\mathbf{p}_1 - \mathbf{p}_2\|_\infty = \|f(\mathbf{v}_1) - f(\mathbf{v}_2)\|_\infty \le \max_k \sum_{j=1}^{4} a_{kj} \|\mathbf{v}_1(j) - \mathbf{v}_2(j)\| \le \|\mathbf{v}_1 - \mathbf{v}_2\|_\infty \tag{58}$$

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:

$$\begin{aligned}
\mathbf{p}(r, \theta) &= \mathbf{p}_{\text{center}} + \frac{r}{r_{\text{corner}}} R(\theta)(\mathbf{p}_{\text{corner}} - \mathbf{p}_{\text{center}}) && (59)\\
&= (I - \frac{r}{r_{\text{corner}}} R(\theta))\mathbf{p}_{\text{center}} + \frac{r}{r_{\text{corner}}} R(\theta)\mathbf{p}_{\text{corner}} && (60)
\end{aligned}$$

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 $r \le r_{\text{corner}}$, we have:

$$\begin{aligned}
\|\mathbf{p}_1(r, \theta) - \mathbf{p}_2(r, \theta)\|_\infty &\le \left\| (I - \frac{r}{r_{\text{corner}}} R(\theta))(\mathbf{p}_{\text{center},1} - \mathbf{p}_{\text{center},2}) \right\|_\infty && (61)\\
&+ \left\| \frac{r}{r_{\text{corner}}} R(\theta)(\mathbf{p}_{\text{corner},1} - \mathbf{p}_{\text{corner},2}) \right\|_\infty && (62)\\
&\le 2\|\mathbf{p}_{\text{center},1} - \mathbf{p}_{\text{center},2})\|_\infty + \sqrt{2}\|\mathbf{p}_{\text{corner},1} - \mathbf{p}_{\text{corner},2})\|_\infty && (63)\\
&\le (2 + \sqrt{2})\|\mathbf{v}_1 - \mathbf{v}_2\|_\infty && (64)
\end{aligned}$$

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

$$\|\mathbf{p}_1 - \mathbf{p}_2\|_\infty = \max_{r,\theta} \|\mathbf{p}_1(r, \theta) - \mathbf{p}_2(r, \theta)\|_\infty \le (2 + \sqrt{2})\|\mathbf{v}_1 - \mathbf{v}_2\|_\infty \tag{65}$$

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

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 $\|\hat{\mathbf{p}} - \mathbf{p}\|_\infty \le \alpha r$. 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

$$\|\mathbf{v} - \mathbf{v}^i\|_\infty \leq \frac{\alpha r}{c} \tag{66}$$

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:

$$\|\mathbf{p} - \mathbf{p}^i\|_\infty \leq c\|\mathbf{v} - \mathbf{v}^i\|_\infty \leq \alpha r \tag{67}$$

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

$$\left\lceil \frac{c}{\alpha} \right\rceil^k \leq \lceil c \rceil^k \left\lceil \frac{1}{\alpha} \right\rceil^{k-d} \left\lceil \frac{1}{\alpha} \right\rceil^d \tag{68}$$

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

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

```
INPUT Parameter distances {Δp_m} with Δp_m ≤ Δp_{m+1}.
INPUT Image distances {ΔI_m}.
INPUT Scale r and noise η.
ΔI_m^+ = max_{1≤l≤m} ΔI_l, for i = 1 ... M.
ΔI_m^- = min_{i≤l≤M} ΔI_l, for i = 1 ... M.
for m = 1 to M do
   Find minimal l* = l*(m) so that ΔI_{l*}^- > ΔI_m^+ + 2η.
   if m ≤ l* then
      Store the 4-tuples (α, γ, A, Γ) = (Δp_m, Δp_{l*}, ΔI_m^+, ΔI_{l*}^-)/r.
   end if
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:*

$$\tilde{A}(\alpha) = \{A : \forall m \ \Delta\mathbf{p}_m \leq \alpha \implies \Delta I_m \leq A\} \tag{69}$$

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

$$\tilde{\Gamma}(\gamma) = \{\Gamma : \forall m \ \Delta\mathbf{p}_m \geq \gamma \implies \Delta I_m \geq \Gamma\} \tag{70}$$

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:*

$$\Delta I_m^+ = \min \tilde{A}(\Delta\mathbf{p}_m) \tag{71}$$

$$\Delta I_m^- = \max \tilde{\Gamma}(\Delta\mathbf{p}_m) \tag{72}$$

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 $\alpha \le \gamma$ always gives the globally optimal solution to the following linear programming:*

$$\min \quad \gamma \tag{73}$$
$$\text{s.t.} \quad \Delta I_m \le A \quad \forall \Delta\mathbf{p}_m \le \alpha \quad (or \;\; A \in \tilde{A}(\alpha)) \tag{74}$$
$$\Delta I_m \ge \Gamma \quad \forall \Delta\mathbf{p}_m \ge \gamma \quad (or \;\; \Gamma \in \tilde{\Gamma}(\gamma)) \tag{75}$$
$$A + 2\eta < \Gamma \tag{76}$$

which has at least one feasible solution $(A \to +\infty, \gamma \to -\infty, \Gamma \to -\infty)$ 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. 73). 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. 74 and Eqn. 75 are satisfied. From the construction of Alg. 1, $A + 2\eta < \Gamma$. Thus, the Algorithm 1 gives a feasible solution to Eqn. 73.

(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

$$A' + 2\eta < \Gamma' \le \Delta I_{l'}^- = \max \tilde{\Gamma}(\gamma') \le \max \tilde{\Gamma}(\gamma) = \Delta I_{l^*}^- \tag{77}$$

On the other hand, $A = \Delta I_m^+ = \min \tilde{A}(\alpha) \le A' \in \tilde{A}(\alpha)$. 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 $\alpha \le \gamma$) gives the optimal pair (α, γ) that satisfies the Relaxed Lipschitz Conditions (Eqn. 20 and Eqn. 21).

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.

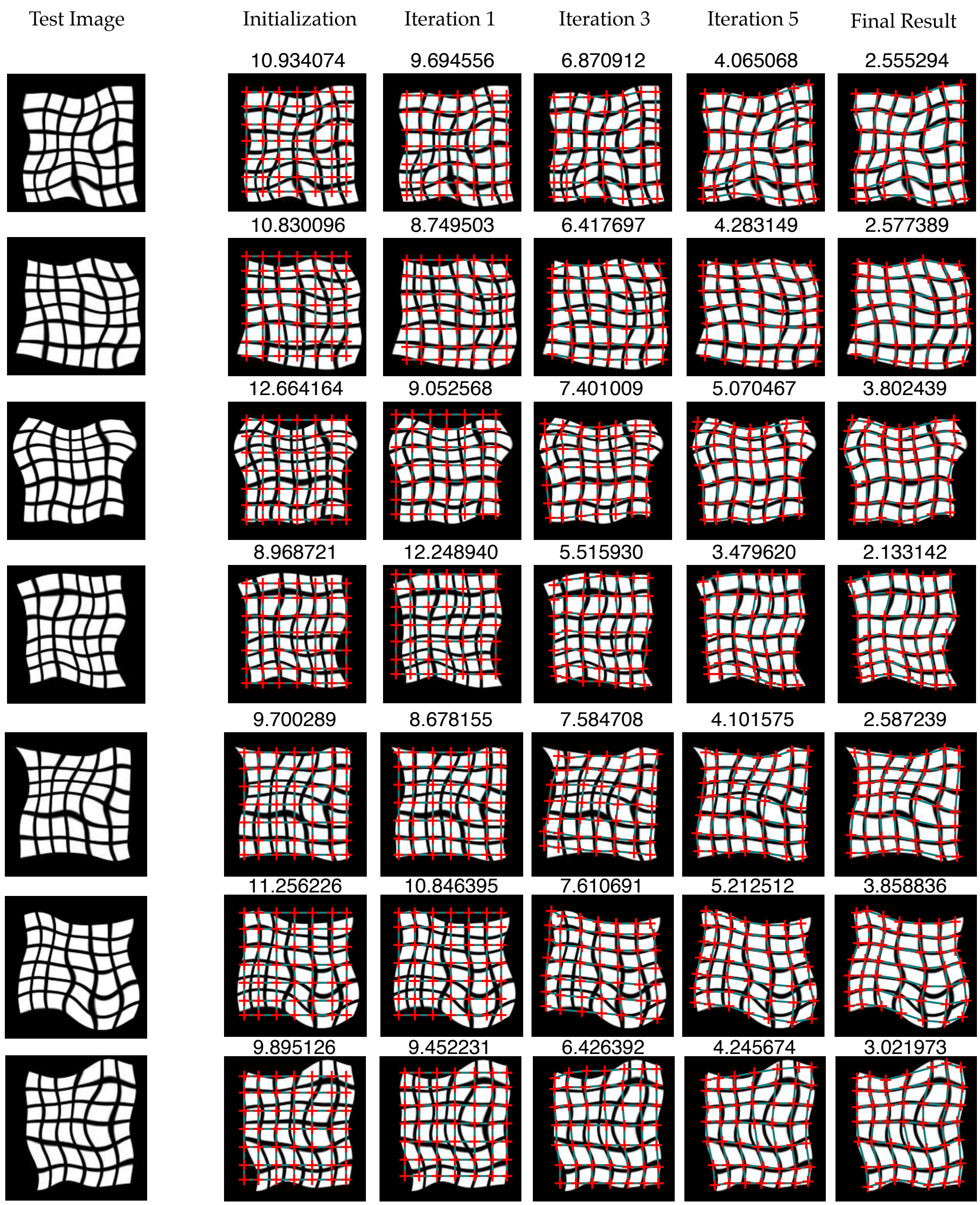


Figure 4: Landmark Estimation at different iterations given by our approach.

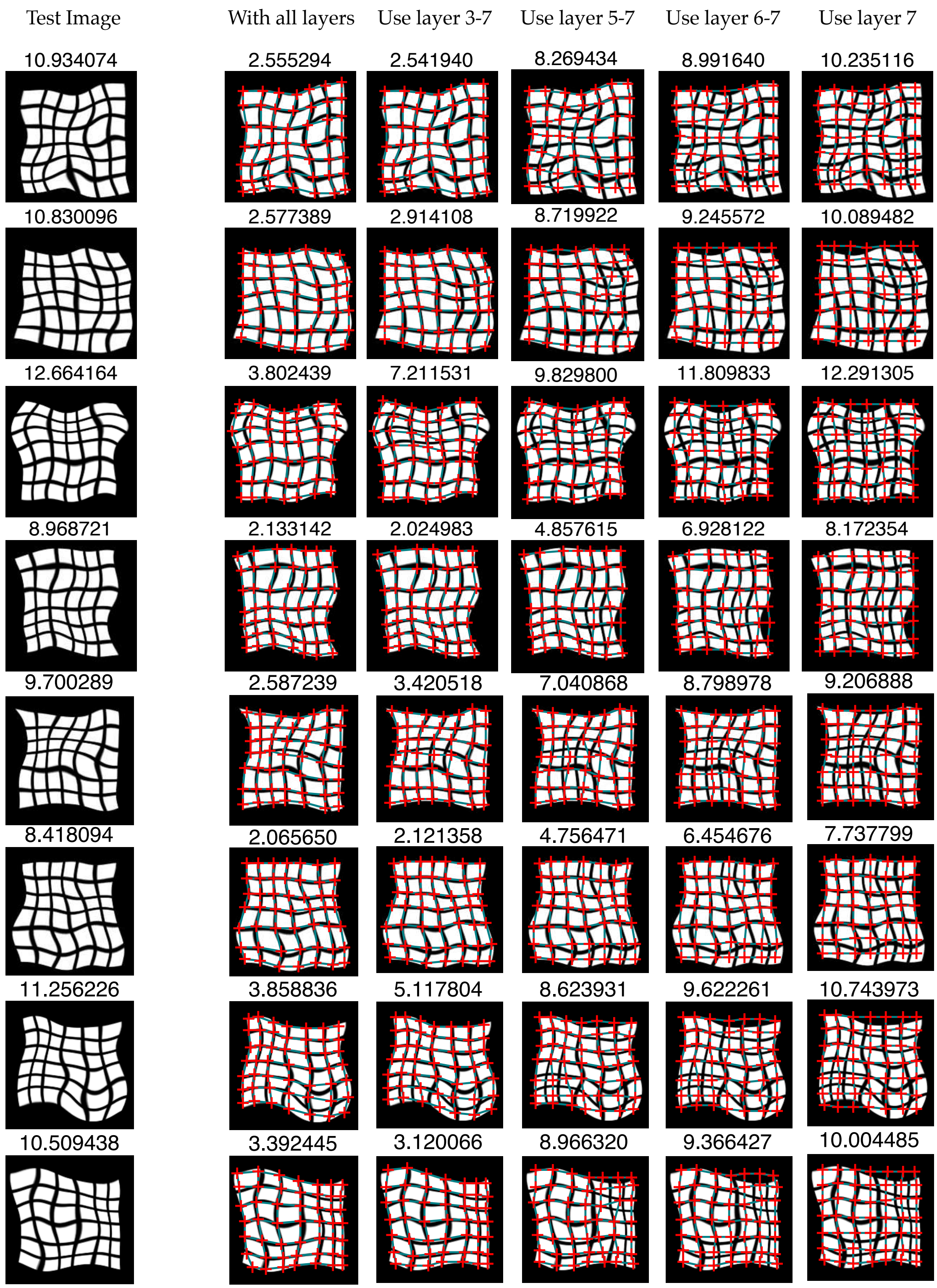


Figure 5: Landmark Estimation using only last L layers of the hierarchy. Layer 3-4 is critical for getting a good estimation of the landmarks on the synthetic data.