Supplementary Material for ProAlignNet : Unsupervised Learning for Progressively Aligning Noisy Contours

VSR Veeravasarapu, Abhishek Goel, Deepak Mittal, Maneesh Singh
Verisk AI, Verisk Analytics
{s.veeravasarapu, a.goel, d.mittal, m.singh}@verisk.com

A1: Proof for $dt[S(\theta_{S\to T})].T = dt[S].T(\theta_{T\to S})$

Under the assumption of infinite boundaryless coordinate systems, $\mathbb{S}$ is source coordinate system. $\theta_{S\to T}$ is a transformation require to align $\mathbb{S}$ with target coordinate system: $\mathbb{T} \equiv \mathbb{S}(\theta_{S\to T})$

Homeomorphism: Most of the transformations we use for contour alignment applications are homeomorphic. The property of homeomorphism states that for each transformation that aligns S to T, there exists an inverse/backward transform that aligns T to S.

$$if\ S(\theta_{S\to T}) \equiv T,\quad \exists\ \theta_{T\to S} \in \Theta\ s.t.\ T(\theta_{T\to S}) \equiv S \quad (1)$$

s and t represent all nonzero pixel locations in S and T respectively.

$$\begin{aligned} dt[S(\theta_{S\to T})].T &= \sum_{t\in\mathbb{T}} \min_{s\in\mathbb{S}(\theta_{S\to T})} \|t-s\|^2 \\ &\quad change-of-coordinates \\ &= \sum_{t\in\mathbb{T}(\theta_{\mathbb{T}\to\mathbb{S}})} \min_{s\in\mathbb{S}} \|t-s\|^2 \\ &= dt[S].T(\theta_{T\to S}) \end{aligned} \quad (2)$$

A2: Min-Max Inequality

Proof for $\min_x \big(f(x) + g(x)\big) \leq \min_x f(x) + \max_x g(x)$

Let us start with below inequality that holds for any x,

$$f(x) + g(x) \leq f(x) + \max_x g(x) \quad (3)$$

The above inequality holds for even smallest possible values of both LHS and RHS.

$$\min_x \big(f(x) + g(x)\big) \leq \min_x \big(f(x) + \max_x g(x)\big) \quad (4)$$

$\max_x g(x)$ is constant wrt x.

$$\min_x \big(f(x) + g(x)\big) \leq \min_x f(x) + \max_x g(x) \quad (5)$$

A3: Detailed derivation for Local shape dependent Chamfer Upperbound

We use min-max inequality to reformulate Eq 5 (in the main paper) so that one can use the concepts of MDT and reparameterization as in Eq 3 and 4 (in the main paper). Min-max inequality states that minimum of the sum of any two arbitrary functions $f(x)$ and $g(x)$ is upper-bounded by the sum of minimum and maximum of individual functions. (Please refer to the appendix A2 for the proofs).

$$\min_x \big(f(x) + g(x)\big) \leq \min_x f(x) + \max_x g(x) \quad (6)$$

Under mild assumptions ($\max_x f \geq \max_x g$), one can prove that this is the tightest possible upperbound. Using the above inequality for the first term on RHS (right hand side) of Eq 5 results in,

$$\begin{aligned} &\sum_{x\in X} \min_{y\in Y} \big(E(x,y) + \lambda E(I'_x, I'_y)\big) \\ &\qquad \leq \sum_{x\in X} \min_{y\in Y} E(x,y) + \lambda \sum_{x\in X} \max_{y\in Y} E(I'_x, I'_y) \end{aligned} \quad (7)$$

Using the inequality for both terms on RHS of Eq 5 results in an upperbound with original Chamfer distance (Eq 2) and shape-dependent terms as follows,

$$\begin{aligned} &C_d(X,Y) \leq C(X,Y) \\ &+ \lambda\left(\frac{1}{N_X}\sum_{x\in X} \max_{y\in Y} E(I'_x, I'_y) + \frac{1}{N_Y}\sum_{y\in Y} \max_{x\in X} E(I'_y, I'_x)\right) \end{aligned} \quad (8)$$

Rewriting the above upperbound in the current context of source to target alignment,

$$\begin{aligned} &C_d(S(\theta),T) \leq C(S(\theta),T) \\ &+ \lambda\left(\frac{1}{N_{S(\theta)}}\sum_{x\in S(\theta)} \max_{y\in T} E(I'_x, I'_y) + \frac{1}{N_T}\sum_{y\in T} \max_{x\in S(\theta)} E(I'_y, I'_x)\right) \end{aligned} \quad (9)$$

We denote this upperbound as C_{up}. As one can observe, the shape-dependent terms are computationally heavy as the maximum being taken over the window of the entire image for each pixel in the other image. However, we can constraint this window to be local and search in the neighborhood defined by that window. Moreover, this maximum-finding operation can be implemented with *MaxPool* layers. Finally, this local shape-dependent Chamfer upperbound is given by,

$$C_{up}(S,T) = \Big(dt[S].T(\theta_{T\to S}) + S(\theta_{S\to T}).dt[T] \Big) + \lambda \Big(\sum_{x\in S(\theta)} \max_{y\in T_x} E(I'_x, I'_y) + \sum_{y\in T} \max_{x\in S_y(\theta)} E(I'_y, I'_x) \Big) \tag{10}$$

As mentioned, this upperbound is a weighted combination of Chamfer loss that measures proximity and a local shape-dependent loss. It is interesting take a closer look at the shape-dependent terms which are distances between two unit gradients.

$$\begin{aligned} E(I'_x, I'_y) &= \sqrt{|I'_x - I'_y|_2^2} \\ &= \sqrt{I'^2_x + I'^2_y - 2I'^T_x.I'_y} \\ &\propto \sqrt{1 - I'^T_x.I'_y} \end{aligned} \tag{11}$$

When the local window is restricted to be 1×1, minimizing the above term is related to maximizing cross-correlation in intensity gradient space. When raw pixel intensities are used in place of gradients, this is maximizing NCC-related metric.

Now the upperbound loss with unit gradients as local shape measures,

$$\begin{aligned} C_{up}\big(S(\theta),T\big) = &\Big(dt[S].T(\theta_{T\to S}) + S(\theta_{S\to T}).dt[T] \Big) \\ &+ \lambda \Big(\sum_{x\in S(\theta)} \max_{y\in T_x} \sqrt{1 - I'^T_x.I'_y} \\ &+ \sum_{y\in T} \max_{x\in S_x(\theta)} \sqrt{1 - I'^T_y.I'_x} \Big) \end{aligned} \tag{12}$$

A4: Background for Multiscale Feature based Approaches

Our designs in this work are majorly inspired by several principles followed in the classical literature. One of them is using multiscale features for incremental alignment. In the classical literature of image registration and optical flow fields, large scale displacements are addressed with multi-scale approaches [9, 1, 10]. These schemes help to increase search scope and escape local minima [3].

These multiscale problem-solving practices have recently regained attention in deep network architectures. Several recent approaches use input images or their features at different resolutions to improve object detection [2] and segmentation [8, 5]. Recent versions of several popular object detection frameworks such as YOLO [7] employed multiscale processes to detect objects at multiple scales of the feature hierarchy independently and fuse them for robust detections. For semantic label and contour estimation, several works [8, 11] exploit lateral/skip connections that associate low-level feature maps across resolutions and semantic levels. Inspired by the effectiveness of multiscale processes in classical and modern literature, we design a deep network that solves alignment problem incrementally at multiple scales.

A5: Background for Progressive Transformations

In classical literature [4, 6] another commonly used practice when estimating complex transformations is to start by estimating a simple transform such as affine and then progressively increase the transform-complexity to refine the estimates along the way. The motivation behind this practice is that estimating a very complex transformation could be hard and computationally inefficient in the presence of noise, so a robust and fast rough estimate of a simpler transformation can be used as a starting point, also regularizing the subsequent estimations of the more complex transformations.

In this work, we follow this practice as we deal with complex misalignments. We start at coarser scales by estimating an affine transformation, which is a linear transformation with 6 degrees-of-freedom (DOF), capable of modeling translation, rotation, non-isotropic scaling, and shear. This estimated affine grid is refined through finer scale networks which employ more flexible transformations (for example, thin-plate splines in this work).

References

[1] Stefan Klein, Marius Staring, Keelin Murphy, Max A Viergever, and Josien PW Pluim. Elastix: a toolbox for intensity-based medical image registration. *IEEE transactions on medical imaging*, 29(1):196–205, 2009. 2

[2] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 2

[3] Tony Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1-2):225–270, 1994. 2

[4] David G Lowe et al. Object recognition from local scale-invariant features. In *iccv*, volume 99, pages 1150–1157, 1999. 2

[5] Tianxiang Pan, Bin Wang, Guiguang Ding, and Jun-Hai Yong. Fully convolutional neural networks with full-scale-features for semantic segmentation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[6] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[7] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 2

[9] Siddharth Saxena and Rajeev Kumar Singh. A survey of recent and classical image registration methods. *International journal of signal processing, image processing and pattern recognition*, 7(4):167–176, 2014. 2

[10] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2432–2439. IEEE, 2010. 2

[11] Zhiding Yu, Chen Feng, Ming-Yu Liu, and Srikumar Ramalingam. Casenet: Deep category-aware semantic edge detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2