Optimal Weights in Convex Functionals: Taking the Guesswork Out of Segmentation

Chris McIntosh and Ghassan Hamarneh

Department of Computer Science, Simon Fraser University, BC, Canada
{cmcintos,hamarneh}@cs.sfu.ca

Abstract. Energy functional minimization is a popular technique for medical image segmentation. The segmentation must be initialized, weights for competing terms of an energy functional must be tuned, and the functional minimized. There is a substantial amount of guesswork involved. We reduce this guesswork by: first, analytically determining optimal weights; second, minimizing a convex energy functional independent of the initialization; and, third, providing an upper bound on the accuracy achievable by a given, convex energy functional. We demonstrate improved results over state of the art on a set of 470 clinical examples.

1 Introduction

Image segmentation remains a daunting task. Many segmentation approaches rely on the minimization of objective functions, including several landmark papers: from the seminal paper of Snakes for 2D segmentation [1] and other explicit models [2] to implicit models [3, 4], graph approaches [5, 6], and variants thereof.

These methods commonly require five building blocks: (*i*) an objective function whose minima provide good segmentations; (*ii*) an appropriate shape representation; (*iii*) a set of parameters including weights to balance the competing terms of the energy functional; (*iv*) an initialization; and (*v*) a method for minimization, whether it be local or global, continuous or combinatorial.

It is in their commonality that the problems lie. The parameter setting, initialization, and minimization phases are well known to be problematic. Often there are unanswered questions: what if a different initialization was used, what about a different minimizer, what if different weights were used between competing terms of the energy functional? Hence, erroneous segmentations cannot be directly attributed to the energy functional or one of the many unknowns of the segmentation process. The solution is to eliminate these points of uncertainty.

Uncertainty with initializations and minimizers can be removed, by formulating problems as convex energy functionals over convex domains. Papers dealing with this issue are now common ground for both continuous [7, 8] and discrete [5, 6] optimization, but they all have uncertainty stemming from the free weights in their energy functionals.

Uncertainty with energy functional weights can be removed by determining the optimal weights for each image to be segmented, else another set of weights

may exist that provides better results. Recently, McIntosh and Hamarneh developed an analytical expression for the optimal functional weights [9]. Their method solves for the optimal functional weights for a training set of image-segmentation pairs, and then infers the optimal parameters for a novel image via geodesic interpolation over the training set. However, their method is not without its drawbacks.

We build upon and extend the work of McIntosh and Hamarneh by addressing two key issues. First, they build their work using non-convex functionals. There is uncertainty that a different initialization or optimization process may have yielded improved results. Instead, in this work, we focus on convex functionals, ensuring global optima and thus removing uncertainty related to local minima and initializations. Second, to evaluate the goodness of weights they formulated an objective function that *itself* includes an implicit weighting between its two competing terms; a weighting which they did not address. We remove the implicit weighting using a convex quadratic formulation under a linear constraint, and thus remove the uncertainty implied by it. Our results show significantly improved accuracy, with a higher degree of certainty.

In addition to the two novel improvements over [9] mentioned above, we emphasize a third contribution. As a by-product of the certainty provided our method, we are now in the position to conduct experiments on training data that measure what we term the *residual error* (RE) of a functional, i.e. the error remaining after all guesswork in the segmentation process has been made optimal. This error cannot be lowered via changes to the initialization, the functional's weights, nor to the minimization process, since they have all been optimized. RE exists because no valid set of weights could be found that forced the correct segmentation to be a minimum of the functional (see sec. 2.3 for details). In this sense, RE is a lower bound on the error obtainable with the chosen energy functional. It is a measure of how good a particular convex energy functional is for segmentation of a data set, as opposed to how well it was tuned; a very important distinction.

2 Theory: Notations and uncertainty in segmentation

In order to more formally explain where the uncertainty lies, and how it needs to be addressed, we first give a more detailed view on the energy minimization based segmentation process. We define a gray-level image I, and its corresponding segmentation S. Then $\mathbf{I}= \{I_1, I_2, ..., I_{\mathbb{N}}\}$ and $\mathbf{S}= \{S_1, S_2, ..., S_{\mathbb{N}}\}$ are training sets of images and their corresponding, correct segmentations.

The first step is the identification of the form of the energy functional. It may be convex or non-convex, as can the shape space over which it is minimized. A common general form is $E(S, I, \mathbf{w}) = w_1 \times internal(S) + w_2 \times external(S, I)$. Notice the free parameter $\mathbf{w} = [w_1, w_2]$. Depending on its value, minima of E favor the internal energy, or the external energy.

The segmentation problem is to solve $S^* = \arg\min_S E(S|I, \mathbf{w})$, which involves choosing a $\mathbf{w}$ and, depending on the nature of the energy functional, it may also

require training appearance and/or shape priors, and setting an initialization. A gradient descent-based solver is typically used but combinatorial approaches have also been explored for discretized versions of the problem [5]. Here we focus on continuous problems, and thus assume a gradient descent solver.

When using gradient descent, non-convexity can be quite problematic. There is no guarantee that another solution does not exist which better minimizes the energy, and thus is potentially a better segmentation. Ideally both functional and shape space are convex; guaranteeing globally optimal solutions.

Simply obtaining a global optima does not, however, guarantee a correct segmentation in the general case. If not appropriately set, the weights $\mathbf{w}$ can cause significant error. Optimizing the weights has been shown to have dramatic effects; reducing error in large data sets by as much as 30% [9]. However, optimizing the weights by hand for even a single image can be a long and tedious task, with no real guarantee of obtaining the correct segmentation.

Instead of *guessing* the optimal weights, suppose we write a function $\gamma(\mathbf{w}|I_j, S_j)$ evaluating how well weight $\mathbf{w}$ works for a given image-segmentation pair (I_j, S_j); such that a parameter is deemed better when it causes S^* to approach S_j, i.e. the minimum of E to be the correct segmentation. Given S_j, we could then calculate the ideal weights for a particular image I_j by solving $\mathbf{w}^* = \arg\min_{\mathbf{w}} \gamma(\mathbf{w}|I_j, S_j)$. It is important that γ itself be convex or globally solvable in $\mathbf{w}$. If γ was not globally solvable, uncertainty would remain in that another $\mathbf{w}^*$ may better minimize γ, and thus better segment the image. Similarly, γ can not contain free parameters, else those parameters would themselves introduce uncertainty; as was the case in [9].

2.1 Convex Energy Functionals

We make use of recent research into convex functionals for image segmentation, specifically that of Cremers et al. where a convex energy functional E, is minimized over a convex shape space represented as probability maps, i.e. $S(x) \in [0, 1]$ for all points x in the image domain Ω, to yield a convex segmentation problem [10]. A shape model is then constructed via principal component analysis (PCA) on a set of training shapes forming a k-dimensional approximation to the shape space, with $\alpha_1, ..., \alpha_k$ eigen coefficients, a mean shape $\bar{S}$, and eigenvectors $\psi_1, ..., \psi_k$. Shapes can now be reconstructed as $S = \bar{S} + \sum_{i=1}^{k} \alpha_i \psi_i$. Writing S in terms of the vector of shape parameters $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^{k}$, convex E can be written as a sum of convex terms:

$$E(\boldsymbol{\alpha}|I = I_j, \mathbf{w} = \widehat{\mathbf{w}}) = \int_{\Omega} \widehat{w}_1 J_1(\boldsymbol{\alpha}|I_j) + ... + \widehat{w}_n J_n(\boldsymbol{\alpha}|I_j) d\mathbf{x} \tag{1}$$

for a fixed image I_j and fixed parameters $\widehat{\mathbf{w}}$, where J_i is a convex term, $\mathbf{w} = [w_1, ..., w_n]$ with $w_i \in [0, 1]$ are weights, and $\widehat{\mathbf{w}}$ is a vector of arbitrarily chosen weights. E is then by definition a convex functional, since the positively weighted

sum of a set of convex terms is itself convex. For proofs of convexity and more details see [10].

Minimizing E optimally can then be performed via gradient descent on $\boldsymbol{\alpha}$ using derivative: $E_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}|I = I_j, \mathbf{w} = \widehat{\mathbf{w}}) = \widehat{w}_1 T_1(\boldsymbol{\alpha}|I_j) + ... + \widehat{w}_n T_n(\boldsymbol{\alpha}|I_j)$ where $E_{\boldsymbol{\alpha}}$ denotes the derivative of E with respect to $\boldsymbol{\alpha}$,and T_i is the derivative of J_ith term. However, since $\widehat{\mathbf{w}}$ is arbitrary nothing can be said about its optimality for the particular image I_j.

2.2 Optimal Energy Functional Weights

For each (I_j, S_j), $I_j \in \mathbf{I}$ and $S_j \in \mathbf{S}$, the task is to find the optimal values for the free weights $\mathbf{w}(I_j)$. This section explores the notion of 'optimal'.

One computationally intractable approach for finding $\mathbf{w}^*$ is to try all possible weight combinations and run the segmentation method then select the weights with the least segmentation error. A better approach, as outlined in [9], is to find the weights $\mathbf{w}^*$ that minimize the magnitude of the derivative, in our case $E_{\boldsymbol{\alpha}}$, of the energy functional at the correct segmentation $\boldsymbol{\alpha}^j$ (i.e. $\boldsymbol{\alpha}^j = [\{\psi\}_{i=1}^k]^+(S_j - \bar{S})$)[1]. Doing so encourages $\boldsymbol{\alpha}^j$ to be a minimum of E (i.e. $E_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^j|I_j, \mathbf{w}^*) = 0$). Since $E_{\boldsymbol{\alpha}}$ is in our case a vector of length k and $w_i(I_j)$ a scalar function, we measure its magnitude as $\left|E_{\boldsymbol{\alpha}}(\mathbf{w}|\boldsymbol{\alpha}^j, I_j)\right|^2$. McIntosh and Hamarneh go further to minimize $|E_{\boldsymbol{\alpha}}(\mathbf{w}|\boldsymbol{\alpha}, I_j)|^2$ for $\boldsymbol{\alpha} = \boldsymbol{\alpha}^j$ while maximizing it for all other possible shapes (in a direction toward the optimal solution). Adopting their approach, for the time being, but with the new convex setup, we proceed as follows.

For a given shape $\boldsymbol{\alpha}^i$, a vector $(\boldsymbol{\alpha}^i - \boldsymbol{\alpha}^j)$ in $\mathbb{R}^k$ represents the direction towards $\boldsymbol{\alpha}^j$. Since $E_{\boldsymbol{\alpha}}(\mathbf{w}|\boldsymbol{\alpha}^i, I_j)$ is the vector in $\mathbb{R}^k$ dictating in what direction, and by what amount, the solution will change at the point $\boldsymbol{\alpha}^i$, a normalized dot-product (projection-like approach) will measure how much in the right direction $E_{\boldsymbol{\alpha}}(\mathbf{w}|\boldsymbol{\alpha}^i, I_j)$ points.

So for an energy functional with a form like those in (1), and following [9], for now, we define $\gamma(\mathbf{w}|I_j, S_j)$ as

$$\gamma(\mathbf{w}) = \left(\left|E_{\boldsymbol{\alpha}}(\mathbf{w}|\boldsymbol{\alpha}^j, I_j)\right|^2 - \lambda \sum_{i \in \mathcal{N}_{\mathbf{S}}} F_{\mathcal{N}_{\mathbf{S}}}(\boldsymbol{\alpha}^j, \boldsymbol{\alpha}^i) \frac{E_{\boldsymbol{\alpha}}(\mathbf{w}|\boldsymbol{\alpha}^i, I_j) \cdot (\boldsymbol{\alpha}^i - \boldsymbol{\alpha}^j)}{|\boldsymbol{\alpha}^i - \boldsymbol{\alpha}^j|} \right) \quad (2)$$

where $\mathcal{N}_{\mathbf{S}}$ denotes a set of nearby (or similar) shapes in the domain of E, and $F_{\mathcal{N}_{\mathbf{S}}}$ is used to weight closer segmentations according to their proximity. The neighborhood $\mathcal{N}_{\mathbf{S}}$ is used instead of the entire shape space to reduce computational complexity. The second term, dubbed the neighborhood term, is negative and $|E_{\boldsymbol{\alpha}}(\mathbf{w}|\boldsymbol{\alpha}^i, I_j)|$ is omitted from the normalized dot-product to reward large steps in the correct direction. Solving $\mathbf{w}^*(I_j) = \arg\min_{\mathbf{w}} \gamma(\mathbf{w}|I_j, S_j)$, yields the optimal weights for image I_j.

[1] We assume the chosen eigenvectors explain 99% of the variance and thus the error incured by representing S as $\boldsymbol{\alpha}$ is negligible.

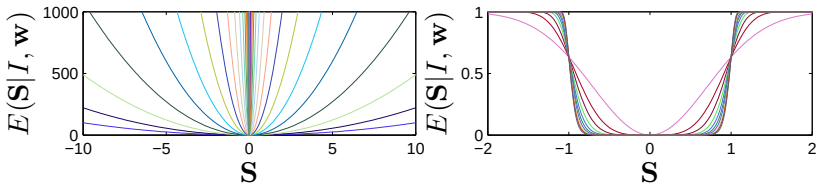

Fig. 1. Varying the shape of energy functionals. Left: Various functionals that show an increasing extent to which the gradient at neighboring shapes points towards the correct segmentation, represented by $S = 0$. Right: Various functionals that show how a set of neighboring shapes can also become minima, a degenerate case.

However, notice that there is a weighting λ between competing terms of (2), which was implicit in [9], and was thus assumed equal to unity and not addressed. With two competing terms, a balance must be struck between: (i) the degree by which $\boldsymbol{\alpha}^j$ is a minimum of E; and (ii) the degree by which the derivative at neighboring points in the shape space points towards $\boldsymbol{\alpha}^j$ (Fig. 1-left). Make λ too small and $\boldsymbol{\alpha}^j$ might be a minimum, but so might the entire neighborhood (Fig. 1-right). Make λ to large and the neighborhood will point in the right direction, but $\boldsymbol{\alpha}^j$ might no longer be a minimum.

To rectify this problem, we make the following observation. By definition, when the energy functional E is convex our only concern is making $\boldsymbol{\alpha}^j$ as much a minimum as possible, while avoiding the degenerate case that the neighboring points are minima (Fig. 1-right). As a result we can move the neighborhood term to a *constraint rather than a cost term* since the degree to which the neighbors point towards $\boldsymbol{\alpha}^j$ does not change whether or not $E_{\boldsymbol{\alpha}}(\mathbf{w}|\boldsymbol{\alpha}^j, I_j) = 0$ (i.e. we must avoid forcing the gradient in the neighborhood to point at $\boldsymbol{\alpha}^j$ at the cost of making $E_{\boldsymbol{\alpha}}(\mathbf{w}|\boldsymbol{\alpha}^j, I_j) \neq 0$). Thus instead of (2) we re-define $\mathbf{w}^*(I_j)$ as

$$\begin{aligned} \mathbf{w}^* &= \arg\min_{\mathbf{w}} \gamma(\mathbf{w}) = \arg\min_{\mathbf{w}} \left| E_{\boldsymbol{\alpha}}(\mathbf{w}|\boldsymbol{\alpha}^j, I_j) \right|^2 \\ \text{s.t.} &\sum_{i \in \mathcal{N}_{\mathbf{S}}} F_{\mathcal{N}_{\mathbf{S}}}(\boldsymbol{\alpha}^j, \boldsymbol{\alpha}^i) E_{\boldsymbol{\alpha}}(\mathbf{w}|\boldsymbol{\alpha}^i, I_j) \cdot \tfrac{(\boldsymbol{\alpha}^j - \boldsymbol{\alpha}^i)}{|\boldsymbol{\alpha}^j - \boldsymbol{\alpha}^i|} > 0 \end{aligned} \tag{3}$$

The result is a convex function in $\mathbf{w}$ under a linear constraint, since the first term in (2) has been shown to be a convex quadratic and the second term linear [9]. A convex function under a linear constraint can be solved via convex optimization, and thus the optimal $\mathbf{w}(I_j)$ is guaranteed.

2.3 Residual Error: An intuitive lower bound for an energy functional

Of interest is that γ, as it appears in (3), allows us to build an intuitive metric for evaluating convex energy functionals on training data; since we have found the weights that force the functional to "do its best". We define the RE as follows. Given E, I_j, S_j, and $\boldsymbol{\alpha}^j$, we solve (3) for $\mathbf{w}^*(I_j)$, then solve $S^* =$

$\arg\min_S E(S|I_j, \mathbf{w}^*)$ initialized with $\boldsymbol{\alpha}^j$, and then calculate the error between S^* and S_j. The resulting error is then not from the initialization, the hand tuning of functional weights, nor from the minimization process. The left over error exists because no set of weights could be found that forced the correct segmentation to be a minimum, while avoiding the degenerate case (Fig. 1-right). Hence, we call it the RE of the functional as it is a measure of how good a particular energy functional is for segmenting a data set, as opposed to how well it was tuned, initialized, and solved; a very important distinction. It enables us to measure not how well a functional did, but how well a functional can do.

3 Method: Segmenting novel images

Given our set of training image-segmentation pairs we will have $\mathbb{N}$ samples of $\mathbf{w}^*(I)$, from which we can interpolate to find values at new points (i.e. novel images). In order to interpolate, we need a metric for measuring distances between images[2]. The set of images with the shortest distances constitutes the neighboring images, $\mathcal{N}_\mathbf{I}$, and $\mathcal{N}_\mathbf{S}$ are their corresponding correct segmentations.

Following [9], we assume $\mathbf{I}$ is smooth over its domain, the space of a particular class of images (e.g. MRI brain scans of normal adults), and that the mapping from images to segmentations is smooth. In other words, we assume that similar images have similar parameters, and similar segmentations. As such, we use a normalized Gaussian kernel, defined over the image distances, to interpolate both the parameters and initializations. For shape and appearance priors, we limit the training data to $\mathcal{N}_\mathbf{I}$ and $\mathcal{N}_\mathbf{S}$, since we are more confident that the correct shape and appearance information is similar to the training data lying in those neighborhoods.

Manifold learning methods are a special class of nonlinear dimensionality reduction techniques that enable the calculation of geodesic distances between data points. We make use of these techniques to calculate distances between both images, and segmentations. Distances between neighboring segmentations allow us to define $F_{\mathcal{N}_\mathbf{S}}(S_i, S_j)$ as a weighting function for the given neighborhood $\mathcal{N}_\mathbf{S}$, normalized to sum to one (i.e. the weight decreases as a function of distance from the center of the neighborhood).

4 Experiments

We validate our method on a set of 470 256×256 affine registered mid-sagittal MR images, with corresponding expert-segmented corpora callosa (CC). Our energy functional takes the form:

$$\begin{aligned} E(\alpha) = &\int_\Omega (w_1(I)f(x)S(x) + w_2(I)g(x)(1 - S(x)) \\ &+ w_3(I)h(x)\,|\nabla S(x)|)\,dx + w_4(I)\alpha^T \Sigma^{-1}\alpha \end{aligned} \tag{4}$$

[2] The choice of metric is beyond the scope of this paper (Sec. 5).

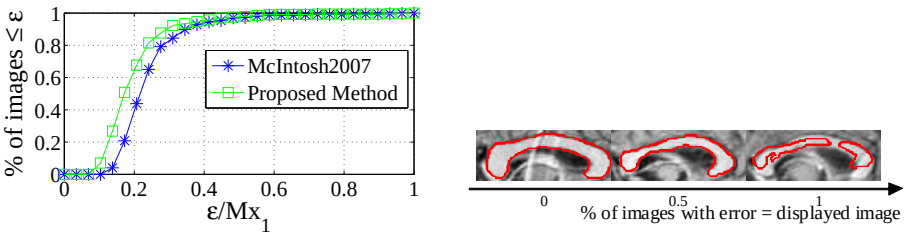

Fig. 2. CC segmentation results. (Left) Error plot where Mx_1 is the maximum measured value for ε. (Right) Segmentations demonstrating the full range of error.

where $f = -log(P_{obj}(I))$, $g = -log(P_{bk}(I))$, for object and background histograms P_{obj}, P_{bk}, $h = \frac{1}{1+|\nabla \mathrm{I}|}$, and Σ^{-1} characterizes the allowable shape distribution (see [10] for details).

To learn the distances, we used a MATLAB implementation of K-ISOMAP [11] from http://isomap.stanford.edu/. As this paper is about the application of optimal parameters to segmentation, issues related to learning the manifold will not be addressed in this work. For K-ISOMAP we set $K = 10$, and reduce the image space to a 5-manifold; chosen as the elbow of the scree plot. For PCA on the shape space we set $k = 10$; the elbow of the corresponding scree plot.

For RE measurements, we perform two different experiments. Firstly, to show that the RE of a perfect functional is indeed zero, we calculate the RE of energy functional (4) plus a fifth term whose unique global minimum is always the correct segmentation for the given image (i.e. we gave the functional weights $\mathbf{w}$ the power to achieve 100% accuracy by adding a rigged, strictly convex term). We measure the error using a modified Dice metric: $\varepsilon = Area(A \cup G - A \cap G)/Area(G)$, where A and G are the binary automatic segmentation and the ground truth, respectively. Using this rigged term, we obtained RE=0 for all 470 images, validating that our method can achieve the full potential of a given functional. Secondly, we calculate the average RE of energy functional (4) on the 470 images as 0.11.

To validate our segmentation method as it pertains to novel data, we perform leave-one-out validation on the set of 470 images using energy functional (4). Under error metric ε, we found the average error to be 0.13, improving over the average error of 0.16 reported in [9]. Our results are summarized in Fig. 2. The figure shows the percentage of images with $\varepsilon \leq$ the specified value on the x-axis. So, for example, with our proposed method approximately 65% of the images have $\varepsilon \leq 0.2$ as opposed to only 44% using the method proposed in [9]. If an error of 0.2 was the cut off for the segmentation method to be clincially useful, our method would have succesfully segmented an additional 21% of the data, or about 100 images more than [9].

5 Discussion

From our experiments, we conclude that the difference $\delta\varepsilon = 0.02$ between what our method achieved ($\varepsilon = 0.13$) and the RE ($\varepsilon = 0.11$) is due to the interpolation over $\mathbf{w}^*(I)$ for novel images, since we know that with optimal weights ε would approach our RE. Based on this result we draw the following conclusion. As opposed to the classical approaches, including [9], that leave us wondering if improved weights (parameters), initializations, or solvers would yield significantly better results, we have presented a method in which we know that in order to improve the results we need a new energy functional (remembering that [9] has an implicit weight in (2) that if adjusted could have improved their results). Consequently, in addition to presenting improved results over a recent technique for an important clinical segmentation task, we have presented a first look at a way to examine and study energy functionals. In doing so, we have taken a step towards reducing the aforementioned guesswork (uncertainty) involved in energy minimization-based segmentation. Our method is not, however, without its drawbacks, as in its application to novel data we introduce an important new question: is this the best image distance metric [12] and interpolation function? Our future work is thus to not only seek improved ways for learning and interpolating over the image domain, but also to find an energy functional with a lower RE. We also would like to formulate the residual error for a more general class of energy functionals than covered by (1), namely a sum of non-convex terms. The approach itself still holds, however the increased complexity of degenerate cases in non-convex functions makes the notion of optimal weights less certain.

References

1. Kass, M., Witkin., A., Terzopoulos, D.: Snakes: Active contour models. IJCV **1(4)** (1987) 321–331
2. Lobregt, S., Viergever, M.: A discrete dynamic contour model. IEEE TMI **14** (1995) 12–24
3. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. In: IJCV. Volume 22(1). (1997) 61–79
4. Chan, T., Vese, L.: Active contours without edges. IEEE TIP (2001) 266–277
5. Boykov, Y., Kolmogorov, V.: Computing geodesics and minimal surfaces via graph cuts. ICCV (2003)
6. Grady, L.: Random walks for image segmentation. IEEE TPAMI **28**(11) (2006) 1768–1783
7. Nikolova, M., Esedoglu, S., Chan, T.F.: Algorithms for finding global minimizers of image segmentation and denoising models. SIAM Journal on Applied Mathematics **66**(5) (2006) 1632–1648
8. Bresson, X., Esedoglu, S., Vandergheynst, P., Thiran, J.P., Osher, S.: Fast global minimization of the active contour/snake model. Journal of Mathematical Imaging and Vision **28**(2) (June 2007) 151–167
9. McIntosh, C., Hamarneh, G.: Is a single energy functional sufficient? adaptive energy functionals and automatic initialization. In: MICCAI (2). (2007) 503–510

10. Cremers, D., Schmidt, F.R., Barthel, F.: Shape priors in variational image segmentation: Convexity, lipschitz continuity and globally optimal solutions. In: IEEE CVPR. (2008)
11. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290** (2000) 2319–23
12. Souvenir, R., Pless, R.: Image distance functions for manifold learning. Image and Vision Computing **25**(3) (2007) 365 – 373 Articulated and Non-rigid motion.