Robust Click-Point Linking for Longitudinal Follow-Up Studies

Kazunori Okada[1] Xiaolei Huang[2] Xiang Zhou[2] Arun Krishnan[2]

[1] Department of Computer Science, San Francisco State University
[2] Computer-Aided Diagnosis and Therapy Solutions, Siemens Medical Solutions
contact: kazokada@sfsu.edu

Abstract. This paper proposes a novel framework for robust click-point linking: efficient localized registration that allows users to interactively prescribe where the accuracy has to be high. Given a user-specified point in one domain, it estimates a single point-wise correspondence between a data domain pair. In order to link visually dissimilar local regions, we propose a new strategy that robustly establishes such a correspondence using only geometrical relations without comparing the local appearances. The solution is formulated as a maximum likelihood (ML) estimation of a spatial likelihood model without an explicit parameter estimation. The likelihood is modeled by a Gaussian mixture whose component describes geometric context of the click-point relative to pre-computed scale-invariant salient-region features. The local ML estimation was efficiently achieved by using variable-bandwidth mean shift. Two transformation classes of pure translation and scaling/translation are considered in this paper. The feasibility of the proposed approach is evaluated with 16 pairs of whole-body CT data, demonstrating the effectiveness.

1 Introduction

This paper presents *robust click-point linking*: a localized registration framework that allows users to interactively prescribe a location where the accuracy has to be high. We assume that a user specifies a point location which is placed near a region of interest in one of the data pair. We call such a user-provided point *point of interest* or *POI*. The task of the interactive localized registration is then to find a single point-wise correspondence: the point in the other data which corresponds to the given POI in the original data. In this study, we consider an application scenario of the longitudinal 3D data studies where a set of follow-up studies of the same patient are subjected for analysis. In this scenario, users may specify a POI by a mouse-click in an arbitrary time-point and mouse cursors for the other time-points are automatically determined as the result of the linking.

One of the main advantages of this approach is that it is faithful to how the registration results are used in practice. In many clinical settings, medical images are only assessed locally. When evaluating a specific lesion or anatomy, the registration accuracy at *the* location must be high. However often practitioners are not concerned if other non-target regions are also correctly registered when they are not looking at them. In comparison to common global registration frameworks, such a local focus of interest also facilitates better *accuracy* and *efficiency* by ignoring influences from, and avoiding computations of, the non-target regions away from a POI. For a standard registration set-up, algorithms are often designed to minimize overall average error. However, such errors, averaged over entire domain, are often hard to interpret by the practitioners in the above clinical context.

On the other hand, the main challenge of this framework is *how to link corresponding regions that are changing or intrinsically different*. Suppose we are to study a follow-up data pair, containing liver tumors imaged before and after

a therapy. For quantifying the therapy's effectiveness, a registration of the data pair would be required, followed by a change analysis. This is a classical circular problem. The registration is required for analyzing interesting temporal changes but the very changes make the registration difficult. The localized registration makes the problem even worse because it demands a harder task of finding a correspondence between visually very dissimilar local regions.

To address the above challenge, we propose a novel linking solution which exploits geometrical contexts of a given POI with respect to pre-computed stable anchor features. The solution also avoids using local appearance-based information that is potentially unreliable. As such anchors, we employ scale-invariant salient-region feature [1–3]. Our approach provides an intuitive geometric formulation of spatial likelihood in a Gaussian mixture form whose maximum likelihood (ML) estimate corresponds to the desired linking solution. We demonstrate that such local ML estimation can be robustly and efficiently solved by using the variable bandwidth mean shift method [4]. This paper presents two instances of the proposed framework for 1) pure translation and 2) scaling and translation. The effectiveness is evaluated by using sixteen whole-body CT follow-up data that are manually annotated.

Our work is related to a number of previous studies. The recent development in the part-based object recognition research [5, 6] has inspired our work. Epshtein and Ullman [6] recently proposed an automatic algorithm for detecting semantically equivalent but visually dissimilar object parts. Our proposed solution can be interpreted as a flexible online version of their batch learning-based framework. Our work is built on recently proposed salient-region feature-based registration [2, 3]. To our best knowledge, however, this is the first attempt to apply the mean shift on these features for solving a registration task. The click-point linking concept has been previously explored in some domain-specific cases e.g., lung nodule detection [7]. Our aim is however to solve this problem in a general setting with an emphasis of handling visually dissimilar regions. Finally landmark-based registration [8] is also related to the proposed framework in the sense that both assume user-provided landmarks specifying where the registration must be accurate. However they aim at completely different technical and application goals. The former finds a smooth domain map from given correspondences while the latter estimates a single correspondence given a POI.

2 Robust Click-Point Linking

First we formally define the robust click-point linking problem. Suppose that a pair of image functions are given to be registered and called *reference image* $I_r(\mathbf{x}_r)$ and *floating image* $I_f(\mathbf{x}_f)$ where $\mathbf{x}_r \in \mathbb{R}^3$ and $\mathbf{x}_f \in \mathbb{R}^3$ represent coordinate variables in their respective continuous domains. The pair of the domains are assumed to be implicitly related by an unknown transformation $\mathcal{T}_\theta : \mathbb{R}^3 \mapsto \mathbb{R}^3$ parameterized by θ so that $\mathbf{x}_r \overset{\mathcal{T}_\theta}{\longmapsto} \mathbf{x}_f$.

Now we suppose that an arbitrary click point $\mathbf{c}_r$ is given as a POI in the reference domain $\mathbf{x}_r$. Then the task of *click-point linking* is defined as the estimation of the point $\mathbf{c}_f$ in the floating domain $\mathbf{x}_f$ which corresponds to the POI $\mathbf{c}_r$ in the reference domain. The true solution $\mathbf{c}_f$ can be defined if we know the true domain transformation $\mathcal{T}_\theta$ such that $\mathbf{c}_f = \mathcal{T}_\theta(\mathbf{c}_r)$.

Next we introduce *salient-region features* whose 3D center coordinate is denoted by $\mathbf{p}$. Suppose we compute N_r features for the reference image forming a set $C_r = \{\mathbf{p}_{r1}, .., \mathbf{p}_{rN_r}\}$ and N_f features for the floating image forming a set $C_f = \{\mathbf{p}_{f1}, .., \mathbf{p}_{fN_f}\}$. Then we let $Q = \{\mathbf{q}_1, .., \mathbf{q}_M\}$ denote a set of M corresponding feature pairs constructed from C_r and C_f where $\mathbf{q}_i = (\mathbf{q}_{ri}, \mathbf{q}_{fi})$, $\mathbf{q}_{ri} \in C_r$, $\mathbf{q}_{fi} \in C_f$, and $M < \min(N_r, N_f)$.

The standard registration solutions aim to estimate the domain transformation $\widehat{\mathcal{T}}_{\hat{\theta}}$ by solving an energy minimization problem $\widehat{\theta} = \text{argmin}_{\theta} E(\theta, I_r, I_f)$. Once the domain transformation is estimated correctly, the click-point linking becomes trivial as $\widehat{\mathbf{c}}_f = \widehat{\mathcal{T}}_{\hat{\theta}}(\mathbf{c}_r)$. However, estimating the transformation from noisy data is far from trivial. The estimation accuracy is very sensitive to the errors in correspondences. The iterative solutions also tend to be computationally expensive.

In our approach, the linking problem is solved by directly optimizing a spatial likelihood function over the location variable $\mathbf{x}_f$ without explicitly estimating the domain transformation,

$$\widehat{\mathbf{c}}_f = \text{argmax}_{\mathbf{x}_f} \mathcal{L}(\mathbf{x}_f | \mathbf{c}_r, Q) \tag{1}$$

where $\mathcal{L}(\mathbf{x}_f | \mathbf{c}_r, Q)$ denotes a spatial likelihood function in the domain of the floating image that is conditional to the POI $\mathbf{c}_r$ in the reference image and a set of corresponding features Q. This generic maximum likelihood formulation allows us to exploit the mean shift algorithm which allows computational efficiency and desired robustness against false correspondences. The following describes details of the solution in steps.

2.1 Salient-Region Feature Extraction and Matching

We use the salient-region features [1–3] as anchor points for constructing geometric contexts in 3D CT volumes. Our implementation follows work by [1–3]. The following briefly describes the main concept.

Given a data point $\mathbf{x}$ and a spherical region $\mathcal{R}_{(s,\mathbf{X})}$ of certain scale described by a radius s and centered at $\mathbf{x}$, the feature extraction provides the best scale $S_{\mathbf{X}}$ and its corresponding saliency value $\mathcal{A}(\mathcal{R}_{(S_{\mathbf{X}},\mathbf{X})})$. Such saliency is defined by the following function $\mathcal{A}(\mathcal{R}_{(S_{\mathbf{X}},\mathbf{X})}) = \mathcal{H}(\mathcal{R}_{(S_{\mathbf{X}},\mathbf{X})}) \cdot S_{\mathbf{X}} \cdot \int_{i(s,\mathbf{X})} \left\| \frac{\partial}{\partial s} p(i|\mathcal{R}_{(s,\mathbf{X})}) \big|_{S_{\mathbf{X}}} \right\| di$ where $p(i|\mathcal{R}_{(s,\mathbf{X})})$ denotes the intensity likelihood estimated by Parzen windows with Gaussian kernels, $i(s,\mathbf{x})$ represents the intensity range within $\mathcal{R}_{(s,\mathbf{X})}$, $\mathcal{H}(\mathcal{R}_{(s,\mathbf{X})})$ is entropy of the intensity distribution, and $S_{\mathbf{X}}$ is the best scale given by maximum entropy such that $S_{\mathbf{X}} = \text{argmax}_s \mathcal{H}(\mathcal{R}_{(s,\mathbf{X})})$. A set of N ($N < 100$) globally most salient features (each defined by its center and the best scale) are extracted by using the following procedure.

Feature Extraction:

A1 For each voxel location $\mathbf{x}$, compute the best scale $S_{\mathbf{X}}$ of the region centered at it, and its saliency value $\mathcal{A}(\mathcal{R}_{(S_{\mathbf{X}},\mathbf{X})})$.

A2 Identify the voxels with local maxima in saliency values. Then the salient regions of interest are those that are centered at these voxels and have the best scales.

A3 Among the local maxima salient regions, pick the N most salient ones $\{\mathbf{p}_i\}$ (with highest saliency values) as region features for the CT volume.

The raw 12-bit CT data can be converted to Hounsfield unit (HU) with the offset of -1024 and the slope of 1. In this study, we use an intensity windowing between 30 and 285 HU for suppressing certain types of non-rigid tissues such as fat (-100 to -50 HU) and water (0 HU) in order to stabilize the feature extraction process.

For both $I_r(\mathbf{x}_r)$ and $I_f(\mathbf{x}_f)$, we independently perform the above feature extraction, resulting in a pair of sets C_r and C_f of N_r and N_f features, respectively. Given a POI $\mathbf{c}_r$ in the reference domain $\mathbf{x}_r$, we find a set Q of M corresponding features, by using the following exhaustive search strategy.

Feature Matching:
B1 Select $M < N_r$ features $\{\mathbf{q}_{r1}, .., \mathbf{q}_{rM}\}$ from C_r which are closest to $\mathbf{c}_r$ in terms of Euclidean distance.
B2 For each reference feature $\mathbf{q}_{ri}$,
  B2a Exhaustively compute similarities against the N_f floating domain features $\{\mathbf{p}_{fj}\}$.
  B2b Select the most similar $\mathbf{p}_{fj}$ and set it as $\mathbf{q}_{fi}$.

Similarity functions used in B2a can be either geometry or appearance based and/or a combination of both. For instance, the appearance similarity can be measured by χ^2 distance between a pair of intensity histograms derived from circular regions $\mathcal{R}_{(S_{\mathbf{q}_{ri}}, \mathbf{q}_{ri})}$ and $\mathcal{R}_{(S_{\mathbf{p}_{fj}}, \mathbf{p}_{fj})}$. This very simple matching algorithm is meant to provide only rough results. It is thus likely that Q contains non-negligible number of false correspondences. However its computational complexity is expected to be significantly lower than other complex approaches, allowing us to realize more efficient solution.

2.2 Spatial Likelihood by Modeling Geometric Contexts

We model the target spatial likelihood function $\mathcal{L}(\mathbf{x}_f|\mathbf{c}_r, Q)$ of the link estimate $\mathbf{c}_f$ in the floating domain $\mathbf{x}_f$ as a L-component Gaussian mixture. We consider a generalized form with a set of all K-subsets of Q. Such a set is denoted by $P = \{P_l | l = 1, .., L\}$ where $L = \binom{M}{K}$ is cardinality of P, $P_l = \{\mathbf{q}_k | k = 1, .., K\}$ is a K-subset of Q, and $\mathbf{q}_k = (\mathbf{q}_{rk}, \mathbf{q}_{fk}) \in Q$ is the k-th correspondence in P_l.

$$\mathcal{L}(\mathbf{x}_f|\mathbf{c}_r, P) = \sum_{l=1}^{L} p(\mathbf{x}_f|\mathbf{c}_r, P_l) \tag{2}$$

$$p(\mathbf{x}_f|\mathbf{c}_r, P_l) = \mathcal{N}(\mathbf{x}_f; \mathbf{m}_l, \sigma_l^2 \mathbf{I}) \tag{3}$$

$$\mathbf{m}_l = f_t(\mathbf{c}_r, P_l) \tag{4}$$

$$\sigma_l = g_t(\mathbf{c}_r, P_l) \tag{5}$$

where f_t and g_t determine the mean and the width of the i-th Gaussian component as a function of the POI and the l-th K-subset P_l of the neighboring correspondence set Q. The form of f_t and g_t depends on the type of transformation $\mathcal{T}_\theta$, mapping the reference domain to the floating domain. This paper derives their closed-form formulae in $\mathbb{R}^3$ for two transformation classes of i) pure translation and ii) scaling and translation, although their extension to more complex projective transformation is also possible using the same strategy. The following describes the basic concept and derivations.

Let us first assume that true domain transformation $\mathcal{T}_\theta$ is modeled by a certain parameterized transformation class. We choose the value of K such that the correspondences in P_l can sufficiently constrain the full degrees of freedom in the transformation system, similar to the well-known RANSAC setup [9]. Unlike the RANSAC that explicitly and iteratively estimates the parameters, the following simple geometrical interpretation allows us to derive the desired closed-form formulae. Suppose that $\mathbf{c}_r$, P_l and unknown $\mathbf{c}_f$ form a pair of polyhedra with $K + 1$ corresponding vertices $(\mathbf{c}_r, \mathbf{q}_{r1}, .., \mathbf{q}_{rK})$ and $(\mathbf{c}_f, \mathbf{q}_{f1}, .., \mathbf{q}_{fK})$. By construction, the polyhedra must satisfy certain geometric invariances under the assumed class of transformation, resulting in equations with the unknown $\mathbf{c}_f$. A closed-form solution of such equations about $\mathbf{c}_f$ provides a form of f_t for the given transformation class. Intuitively this general procedure can be understood

as to i) define the known geometric configuration of $\mathbf{c}_r$ relative to the context $\{\mathbf{q}_{rk}\}$, ii) transfer such context under the assumed transformation to the floating domain with $\{\mathbf{q}_{fk}\}$, and iii) determine $\mathbf{c}_f$ analytically given $\{\mathbf{q}_{fk}\}$.

First we introduce a pair of local coordinate frames for describing the polyhedra. Let $\mathbf{c}_{rl}$ and $\mathbf{c}_{fl}$ denote position vectors of $\mathbf{c}_r$ and $\mathbf{c}_f$ in the l-th local reference and floating frames whose origin are set at an arbitrary chosen feature $\mathbf{q}_l = (\mathbf{q}_{rl}, \mathbf{q}_{fl})$ from P_l such that $\mathbf{c}_r = \mathbf{c}_{rl} + \mathbf{q}_{rl}$ and $\mathbf{c}_f = \mathbf{c}_{fl} + \mathbf{q}_{fl}$. Then $\mathbf{c}_{fl}$ is the unknown that must be estimated given $\mathbf{c}_r$ and P_l.

When $K = 1$, we have $P = Q$ and $L = M$. This sufficiently constrains only pure translation case. The derivation of f_t is straightforward. Vectors are invariant under the assumed pure translation, resulting in an equation $\mathbf{c}_{fl} = \mathbf{c}_{rl}$. The solution immediately gives

$$\mathbf{m}_{l,K=1} = f_{t,K=1}(\mathbf{c}_r, P_l) = \widehat{\mathbf{c}}_f = \mathbf{c}_r - \mathbf{q}_{rl} + \mathbf{q}_{fl} \tag{6}$$

When $K = 2$, each P_l yields two correspondences providing 6 constraints in $\mathbb{R}^3$. These constraints are sufficient to determine scaling and translation (4 DOF) and pure translation (3 DOF). Let $\mathbf{q}_{la} = (\mathbf{q}_{rla}, \mathbf{q}_{fla})$ denote a single remainder after choosing $\mathbf{q}_l$ from P_l. This results in a pair of similar triangles $(0, \mathbf{q}_{rla} - \mathbf{q}_{rl}, \mathbf{c}_{rl})$ and $(0, \mathbf{q}_{fla} - \mathbf{q}_{fl}, \mathbf{c}_{fl})$ without rotation. Between the pair, therefore, corresponding normalized vectors and ratio of corresponding vector norms are invariant, resulting in $\frac{\mathbf{c}_{fl}}{\|\mathbf{c}_{fl}\|} = \frac{\mathbf{c}_{rl}}{\|\mathbf{c}_{rl}\|}$ and $\frac{\|\mathbf{c}_{fl}\|}{\|\mathbf{q}_{fla} - \mathbf{q}_{fl}\|} = \frac{\|\mathbf{c}_{rl}\|}{\|\mathbf{q}_{rla} - \mathbf{q}_{rl}\|}$ where $\|\cdot\|$ denote a vector norm. After some algebra, the desired function estimating the l-th Gaussian component mean with $K = 2$ is derived as follows.

$$\mathbf{m}_{l,K=2} = f_{t,K=2}(\mathbf{c}_r, P_l) = \widehat{\mathbf{c}}_f = \frac{\|\mathbf{q}_{fla} - \mathbf{q}_{fl}\|}{\|\mathbf{q}_{rla} - \mathbf{q}_{rl}\|}(\mathbf{c}_r - \mathbf{q}_{rl}) + \mathbf{q}_{fl} \tag{7}$$

For modeling the Gaussian width, we can interpret scales $S_{\mathbf{q}_{rk}}$ and $S_{\mathbf{q}_{fk}}$ of the salient-region features in P_l as statistical uncertainty for localizing the feature points. In this paper we assume that deformation due to the domain transformation is not too large, allowing us to ignore the uncertainty propagation factor. Therefore the uncertainties at the features can also be treated as uncertainties at the estimated component mean.

$$\sigma_l = g_t(\mathbf{c}_r, P_l) = \frac{\sum_{k=1}^{K} S_{\mathbf{q}_{rk}} + \sum_{k=1}^{K} S_{\mathbf{q}_{fk}}}{2K} \tag{8}$$

2.3 Mean Shift-based Robust Maximum Likelihood Estimation

This section describes our robust and efficient solution for the maximum likelihood estimation problem in (1) with the likelihood model (2-5). Due to the feature matching errors discussed in Sec. 2.1, the likelihood function becomes multi-modal with the false correspondences creating outlier (largely deviated) modes. Our task becomes estimating the mixture mode due only to the correctly found correspondences. We solve this task by using the variable-bandwidth mean shift (VBMS) proposed in [4]. VBMS is extension of the original mean shift to spatially variant bandwidth case where different data points may have different significance. This extension allows its application to solve an information fusion problem where the task is to estimate the most plausible solution given a set of hypotheses described in a Gaussian mixture model.

Let $\mathbf{x}_i \in \mathbb{R}^3, i = 1, .., M$ denote a set of 3D data points, and H_i is a 3D matrix indicating uncertainty or significance associated with the point $\mathbf{x}_i$. The point density estimator with 3D normal kernel at the point $\mathbf{x}$ is given by $\widehat{f}_v(\mathbf{x}) = \sum_{i=1}^{M} \mathcal{N}(\mathbf{x}; \mathbf{x}_i, H_i) = \frac{(2\pi)^{-3/2}}{M} \sum_{i=1}^{M} |H_i|^{-1/2} \exp(-\frac{1}{2}(\mathbf{x} - \mathbf{x}_i)^T H_i^{-1}(\mathbf{x} - \mathbf{x}_i))$. The VBMS vector $\mathbf{m}_v(\mathbf{x})$ is then defined by

$$\mathbf{m}_v(\mathbf{x}) = H_h(\mathbf{x}) \sum_{i=1}^{M} w_i(x) H_i^{-1} \mathbf{x}_i - \mathbf{x} \tag{9}$$

where $H_h(\mathbf{x})$ denotes the data-weighted harmonic mean of the bandwidth matrices at $\mathbf{x}$ such that $H_h^{-1}(\mathbf{x}) = \sum_{i=1}^{M} w_i(x) H_i^{-1}$. The weight $w_i(x)$ represents the influence from i-th component at $\mathbf{x}$ normalized over all the components $w_i(x) = \frac{|H_i|^{-1/2} \exp(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)^T H_i^{-1}(\mathbf{x}-\mathbf{x}_i))}{\sum_{i=1}^{M} |H_i|^{-1/2} \exp(-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)^T H_i^{-1}(\mathbf{x}-\mathbf{x}_i))}$. It can be shown that the VBMS vector is an adaptive estimator of normalized gradient of the underlying density such that $\mathbf{m}_v(\mathbf{x}) = H_h(\mathbf{x}) \frac{\widehat{\nabla} f_v(\mathbf{x})}{\widehat{f}_v(\mathbf{x})}$. The following iterative algorithm with the VBMS vector is provably convergent to a mode of the density estimate in the vicinity of the initialization $\mathbf{x}_{init}$ in the gradient-ascent sense but without nuisance parameter tuning

$$\begin{aligned} \mathbf{y}_0 &= \mathbf{x}_{init} \\ \mathbf{y}_{n+1} &= \mathbf{m}_v(\mathbf{y}_n) + \mathbf{y}_n \end{aligned} \tag{10}$$

We denote the convergence of the iterator by $\mathbf{y}^*$.

We apply VBMS to our problem by simply setting $\mathbf{x}_i = \mathbf{m}_l$ and $H_i = \sigma_l^2 \mathbf{I}$ as defined in (4) and (5), respectively. Our solution performs a single VBMS iteration from an initialization $\mathbf{x}_{init}$ estimated from C_r and C_f.

Local ML Estimation by VBMS:

C1 Compute the means $\mathbf{z}_r$ and $\mathbf{z}_f$ of salient-region feature points in C_r and C_f, respectively.
C2 Compute the mean bias $\mathbf{z} = \mathbf{z}_f - \mathbf{z}_r$ between C_r and C_f.
C3 Set the initialization of a VBMS iterator by the mean bias-corrected POI in the floating domain: $\mathbf{x}_{init} = \mathbf{c}_r + \mathbf{z}$
C4 Perform the VBMS algorithm in (10), resulting in the convergence $\mathbf{y}^*$.
C5 Results in the linking estimate $\widehat{\mathbf{c}}_f = \mathbf{y}^*$.

3 Experimental Studies

The feasibility of the proposed framework is evaluated by testing the 3D implementation of the above algorithm with a set of 16 whole-body CT volume pairs. Two volumes in each pair are scans taken at different time-points of the same patient. The same scanner protocols were used between each pair. The original volume with a stack of 512-by-512 axial slices are down-sampled to 128-by-128 slices. One of each pair is arbitrary picked to be a reference image, leaving the other to be a floating image.

The following setting of the proposed algorithm was used. For each volume, a number of 50 salient-region features are pre-computed: $N_r = N_f = 50$. For each click-point $\mathbf{c}_r$, the feature matching algorithm is then performed with 10 nearest reference features: $M = 10$. Two similarity functions are considered in this study:

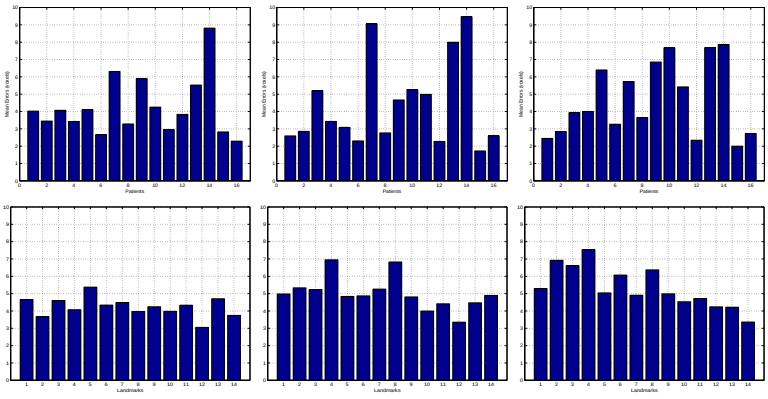

Fig. 1. Experimental results. Top: average errors as a function of 16 different patients. Bottom: average errors as a function of 14 different landmarks. For feature matching, we consider two versions of similarity function. S1: geometric Euclidean distance. S2: unbiased linear combination of the geometric and appearance similarity (χ^2 intensity histogram distance). Left: S1 without template-based correction. Middle: S1 with template-based correction. Right: S2 with template-based correction. All the errors are calculated with the unit of voxels.

geometric Euclidean distances and the χ^2 distance of intensity histograms. Two solutions for 1) pure translation with $K = 1$ and 2) scaling and translation with $K = 2$ are considered. For testing, we used pre-recorded 3D landmarks that are manually labeled by experts. There were 14 landmarks for each person distributed at significant anatomical landmarks, including pelvis, lung, kidneys, and collar bones. For each pair, these 14 points in the reference image are used as POIs and Euclidean errors are computed between the estimated links $\mathbf{c}_f$ and the ground-truth landmarks in the floating domain of $\mathbb{R}^3$. The total of 224 test cases (16 patients over 14 landmarks) were evaluated. We also consider a post-process for refining the estimated click-point by using a template matching-based refinement. The size of the spherical template around each landmark was automatically estimated by using the maximum entropy criterion [2].

Fig. 1 shows the result of our experiments. The top row shows the average errors plotted over different patients. On the other hand, the bottom row shows those plotted over different landmarks. For feature correspondence matching, we consider two versions of similarity function. One was the geometric Euclidean distance with the mean bias adjustment and the other was a linear combination of the geometric distance and an appearance-based distance using χ^2 distance of intensity histograms. The left column shows the results with the geometric Euclidean distance. The total average and median errors were 4.23 and 3.50 voxels, respectively. The middle column shows the results with the geometric distance and the refinement. The average and median errors were 4.39 and 3.24, respectively. Finally, the results with the appearance-based similarity as well as the post-refinement are shown in the right column. The average and median errors were 4.68 and 3.10, respectively. For extracting 50 features in a 3D volume with 128 by 128 slices, it took roughly 2.5 minutes while it took only a fraction of second for the rest of processing.

Overall, the average errors were in the range of 3 to 5 voxels, demonstrating the feasibility of the proposed methods. The results also show that the accuracy

depends strongly on patients but not as strongly on landmarks. Visual inspection revealed that higher errors (e.g, patient 7 and 14) were caused mainly by the outlier failures due to lack of corresponding features between pairs. The usage of the appearance-based similarity and post-refinement slightly improved accuracy. However the improvement was small and made outlier errors actually worse. For the inliers, the average errors were smaller than 3 voxels with the post-refinement.

4 Conclusion and Future Work

This article proposed a novel framework for robust click-point linking. In order to derive a robust solution for linking visually dissimilar local regions, such as changing tumors, we proposed a framework for a mean shift-based ML estimation over a Gaussian mixture likelihood that models geometric context of arbitrary click-points with respect to salient-region features. Our experimental study demonstrated the robustness of the proposed approach using hand-labeled whole-body CT data set. We are currently working on extending our current solutions to account for uncertainty propagation and similarity and affine transformation. We also plan to further improve robustness and efficiency of the salient-region feature extraction and matching parts.

Acknowledgments

We would like to thank Chengyang Xu, Yiyong Sun, Yanghai Tsin, Yakup Genc, Maneesh Singh and Gareth Funka-Lea for their useful comments, fruitful discussions, and continuing supports.

References

1. Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision **45** (2001) 83–105
2. Huang, X., Sun, Y., Metaxas, D., Sauer, F., Xu, C.: Hybrid image registration based on configural matching of scale-invariant salient region features. In: Second IEEE Workshop on Image and Video Registration. (2004)
3. Hahn, D., Sun, Y., Hornegger, J., Xu, C., Wolz, G., Kuwert, T.: A practical salient region feature based 3D multimodality registration method for medical images. In: SPIE Med. Imag. (2006)
4. Comaniciu, D.: An algorithm for data-driven bandwidth selection. IEEE Trans. Pat. Anal. Mach. Intell. **25** (2003) 281–288
5. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: IEEE Conf. on Computer Vision and Pattern Recognition. Volume 2. (2003) 264–271
6. Epshtein, B., Ullman, S.: Identifying semantically equivalent object fragments. In: IEEE Conf. on Computer Vision and Pattern Recognition. Volume 1. (2005) 2–9
7. Novak, C., Shen, H., Odry, B., Ko, J., Naidich, D.: System for automatic detection of lung nodules exhibiting growth. In: SPIE Med. Imag. (2004)
8. Pennec, X., Ayache, N., Thirion, J.: Landmark-based registration using features identied through differential geometry. In: Handbook of Medical Imaging. Academic Press (2000) 499–513
9. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Comm. of the ACM **24** (1981) 381–395