Robust Click-Point Linking: Matching Visually Dissimilar Local Regions

Kazunori Okada
Department of Computer Scienc
San Francisco State University
San Francisco, CA
kazokada@sfsu.edu

Xiaolei Huang
Department of Computer Science and Engineering
Lehigh University
Bethlehem, PA
xih206@lehigh.edu

Abstract

This paper presents robust click-point linking: a novel localized registration framework that allows users to interactively prescribe where the accuracy has to be high. By emphasizing locality and interactivity, our solution is faithful to how the registration results are used in practice. Given a user-specified point, the click-point linking provides a single point-wise correspondence between a data pair. In order to link visually dissimilar local regions, a correspondence is sought by using only geometrical context without comparing the local appearances. Our solution is formulated as a maximum likelihood estimation (MLE) without estimating a domain transformation explicitly. A spatial likelihood of Gaussian mixture form is designed to capture geometrical configurations between the point-of-interest and a hierarchy of global-to-local 3D landmarks that are detected using machine learning and entropy based feature detectors. A closed-form formula is derived to specify each Gaussian component by exploiting geometric invariances under specific group of domain transformation via RANSAC-like random sampling. A mean shift algorithm is applied to robustly and efficiently solve the local MLE problem, replacing the standard consensus step of the RANSAC. Two transformation groups of pure translation and scaling/translation are considered in this paper. We test feasibility of the proposed approach with 16 pairs of whole-body CT data, demonstrating the effectiveness.

1. Introduction

The main motivation of our work is to address clinical demands for *locality* and *interactivity* in image registration, which has not been well-addressed within traditional frameworks. Given a pair of data to be compared, a domain registration is commonly performed for the data pair in order to enable change analyses with necessary domain alignment. Such a registration is often carried out as an offline preprocess due to its high computational complexity. In practice, this high complexity prohibits us from using a registration solution interactively. On the other hand, the standard (rigid) registration algorithms are of global nature, designed to minimize an expert-designed error function which averages/integrates specific local errors over entire domain. Due to this, it is often difficult to predict where accurate registration should occur. Moreover such globally averaged error is hard to interpret toward specific clinical contexts by the practitioners.

In many clinical settings, however, medical images are only assessed locally at times but sequentially in an interactive fashion. When evaluating a specific lesion or anatomy, the registration accuracy at *the* location must be high. Practitioners are also not concerned if other non-target regions are also correctly registered when they are not looking at them. For example, in the longitudinal 3D data studies for cancer therapy monitoring, a set of follow-up studies of a patient with multiple lesions would be analyzed for each lesion and their potential metastases one by one sequentially. These clinical demand and context have not been well addressed and cannot be exploited by the above global rigid registration approach.

Addressing these issues, we propose *robust click-point linking*: a localized registration framework that allows users to interactively prescribe a location where the accuracy has to be high. Suppose that a user/practitioner specifies a 3D data point location near a region of interest in one of the data pair. We call such a user-provided data 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.

This computational framework is designed to be faithful to how the registration results are used for the longitudinal study practice. In this scenario, practitioners 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. In comparison to the common global registration frameworks, the local-

ity emphasis facilitates better *accuracy* and *efficiency* by ignoring influences from, and avoiding computations of, the non-target regions away from a POI. On the other hand, the interactivity emphasis yields a tool with better user-centric accuracy. Such a tool meets the above clinical demands by providing practitioners a control for choosing locations at which accuracy must be high.

The main technical challenge, however, 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, as we propose, makes the problem even harder because it demands a harder task of finding a *correspondence between visually very dissimilar local regions*. This consideration for matching visually dissimilar local regions renders typical solutions ineffective. Template matching [1] offers a natural solution to the click-point linking problem by constructing a visual template centered at a POI and search the other image with it. Although such a solution fits well to our aim, it should obviously fail to match visually dissimilar local regions.

To address the above challenge, we propose a novel linking solution which exploits geometrical context information. We model the geometrical context as a set of relative configurations between a given POI to pre-computed stable anchor feature points. By matching such geometrical model, the proposed solution avoids matching an appearance-based local feature that can be unreliable. The stable anchor feature points are detected in a two-level process. First, using a 3D object detection algorithm that stems from real-time face detection [2], a classifier detector is learned for each among a set of stable whole-body landmarks, based on information from a large set of training volumes. Hence given a reference and a floating volume to match, the learned detectors are applied to extract global whole-body landmarks on both volumes. Second, near the POI, we also generate a number of local anchor points that can be reliably matched between the two volumes. To this end, we employ scale-invariant salient-region feature [3, 4] for detecting the local anchors and exhaustive nearest neighbor search for finding correspondences. In order to robustify against correspondence errors intrinsic to the simple matching technique above, we adopt RANSAC [5] approach. Our novel contribution is to extend the RANSAC to our click-point linking context. Instead of explicitly estimating the underlying domain transform as in the original RANSAC, our approach treats such transform as implicit knowledge and estimates point-wise linking hypothesis directly. Such a direct linking estimator is derived as a closed-form formula by solving a set of equations representing geometric invariances, under specific transformation group and data dimensionality, between a pair of polyhedra. This approach is efficient because it avoids estimating the transform which is unnecessary for our linking problem. The consensus among the multiple linking hypotheses is achieved by using a mean shift algorithm. Together with confidence measures, derived as a function of the saliency scales, the set of multiple linking hypotheses are interpreted as a spatial likelihood in a Gaussian mixture form whose maximum likelihood estimate (MLE) corresponds to the desired linking solution. We demonstrate that such local MLE can be robustly and efficiently solved by using the variable bandwidth mean shift method [6]. 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.

1.1. Related Work

The recent development in the part-based object recognition research [7, 8] has inspired our work. Epshtein and Ullman [8] 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. The click-point linking concept has been previously explored in some domain-specific cases e.g., lung nodule detection [9]. Our aim is however to solve this problem in a general setting with an emphasis of handling visually dissimilar regions. The spatial likelihood formulation with its mean shift solution is our unique contribution in this context. Our idea to exploit the geometric invariance for extending RANSAC toward the mean shift-based MLE problem is new to our best knowledge and provides a generic tool beyond the specific application focus of this article. A previous work presented in [10] employed an entropy-based salient region detector [3] to find anchor features, but the method was sensitive to noise and anchor point correspondence errors in one part of the volume could get propagated to another part. In this paper we detect a hierarchy of both global and local anchor points. The global anchors include a sparse set of whole-body landmarks detected with a machine learning approach, and on top of that, we use salient region detector to extract more salient points as local anchors near the POI. In more traditional setting, non-rigid registration [11, 12, 13] aims to achieve locally accurate global registration by allowing non-rigid transform between the data domain pair. At a visually dissimilar local region, however, this approach tends to be suboptimal because correspondence can be achieved only implicitly via smoothness assumption of neighboring transforma-

tion. Moreover typical iterative solution for this approach tends to be time consuming due to the increased degrees of freedom to be estimated. On the other hand, similar to our approach, feature-based registration estimates a domain transform using a set of points [14] or curves/surface patches [15, 16]. Both of these traditional approaches, however, focus on global domain registration and do not offer simple ways to make the registration process interactive and local. Finally, landmark-based registration [17] is also related to our 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

This section formally introduces the robust click-point linking problem. Suppose that a pair of images are given to be registered. Without loss of generality, we call one *reference image* $I_r(\mathbf{x}_r)$ and the other *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 related by an unknown linear 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 $\mathbf{c}_r \in \mathbb{R}^3$ 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 $\mathbf{c}_f \in \mathbb{R}^3$ 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 $\mathbf{c}_f = \mathcal{T}_\theta(\mathbf{c}_r)$.

The standard registration solutions aim to estimate the domain transformation $\widehat{\mathcal{T}}_{\hat{\theta}}$ by solving a data-driven 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 in the feature-based framework, for example. The iterative solutions can also 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 landmark features Q.

The set Q contains N corresponding *landmark features*, forming the geometrical context of the POI $\mathbf{c}_r$. We use the machine learning and entropy-based approaches to detect a hierarchy of global-to-local landmarks in I_r and I_f as described in the next section, resulting in

$$C_r = \{\mathbf{p}_{r1}, .., \mathbf{p}_{rN}\}, \quad C_f = \{\mathbf{p}_{f1}, .., \mathbf{p}_{fN}\}$$

Then a set of M corresponding feature pairs is constructed from C_r and C_f

$$Q = \{\mathbf{q}_1, .., \mathbf{q}_M\}$$

where $\mathbf{q}_i = (\mathbf{q}_{ri}, \mathbf{q}_{fi})$, $\mathbf{q}_{ri} \in C_r$, $\mathbf{q}_{fi} \in C_f$, and $M \leq N$.

The above generic maximum likelihood formulation allows us to exploit the mean shift algorithm which results in computational efficiency and desired robustness against false correspondences. The following describes details of the solution in steps.

2.1. Global-to-local Landmark Detection and Matching

In order to robustly detect a sufficiently number of anchor landmarks for constructing the geometric context in 3D CT volumes, we use a learning-based object detection approach to detect a sparse set of whole-body landmarks as stable global anchors, and near the POI, we apply an entropy-based salient region detector to extract a denser set of salient points as local anchors.

The whole-body landmark detection algorithm stems from the real-time face detection algorithm [2] in computer vision. It learns a classifier (or detector) for each landmark based on its neighborhood appearances in a large set of training volumes. An intermediate representation, *integral image*, together with Cascaded AdaBoost Training on simple 3D rectangular box features that are reminiscent of Haar basis functions, allow rapid processing of images while achieving high detection rate. For training purposes, we collected $K(= 46)$ whole-body CT volumes from normal subjects, and in each volume, three experts are asked to manually place $N(= 18)$ landmarks, as consistently as possible. The landmarks are distributed in the head, neck, chest, abdomen and pelvis regions. To deal with challenges such as added computational complexity in 3D and the need for reliably detecting multiple targets instead of one target, we train a multi-resolution classifier in the scale space for each landmark, have each classifier output several detection candidates, learn a Point Distribution Model (PDM) [18] to represent the probabilistic spatial distribution of landmarks, and use the PDM model to assist in selecting the top candidate for each landmark so that both local appearance and global context constraints are satisfied.

Let us denote the detected whole-body landmarks on the reference image I_r as $A_r = \{\mathbf{p}_{r1}, .., \mathbf{p}_{rn}\}$, and the landmarks on the floating image I_f as $A_f = \{\mathbf{p}_{f1}, .., \mathbf{p}_{fn}\}$,

then correspondences between A_r and A_f can be established trivially since the identity of each landmark is known. We let $Q_A = \{\mathbf{q}_1, .., \mathbf{q}_n\}$ denote the set of corresponding landmark feature pairs constructed from A_r and A_f, where $\mathbf{q}_i = (\mathbf{q}_{ri}, \mathbf{q}_{fi})$, $\mathbf{q}_{ri} \in A_r$ and $\mathbf{q}_{fi} \in A_f$. These whole-body landmarks and their correspondences provide global geometric contexts in 3D CT volumes, and we also utilize them to achieve a global rough alignment of I_r and I_f using the Least-squares weighted alignment algorithm [19].

We further establish local geometric context near the POI, $\mathbf{c}_r$, using salient region features [3, 4]. Our implementation is similar to that described in [10]. Instead of looking for salient points in the entire image however, we only utilize salient points in a subvolume centered at the POI, and with a radius defined by a few closest whole-body landmarks. First, in the subvolume on I_r, a set of N_r salient region features, B_r, are detected. Second, a corresponding subvolume on I_f is computed based on whole-body landmark correspondences, and in this subvolume, a set of N_f salient region features, B_f, are detected. We then find a set Q_B of corresponding salient region features by using the following exhaustive search strategy.

Local Salient Region Feature Matching:

A1 Select $m < N_r$ features $\{\mathbf{o}_{r1}, .., \mathbf{o}_{rm}\}$ from B_r which are closest to the POI, $\mathbf{c}_r$, in terms of Euclidean distance.

A2 For each reference feature $\mathbf{o}_{ri}$,

- **A2a** Exhaustively compute similarities against the N_f floating domain features $\{\mathbf{o}_{fj}\}$. The similarity functions can be either geometry or appearance based and/or a combination of both.
- **A2b** Select the most similar $\mathbf{o}_{fj}$ and set it as $\mathbf{o}_{fi}$.
- **A2c** Add the correspondence $\mathbf{q}_i = (\mathbf{o}_{ri}, \mathbf{o}_{fi})$ to the set Q_B.

Now putting global whole-body landmarks together with local salient points, we have the complete set of anchor features: $C_r = \{A_r, B_r\}$, $C_f = \{A_f, B_f\}$, and a set of $(n+m)$ corresponding feature pairs: $Q = \{Q_A, Q_B\}$.

The above global-to-local anchor feature detection scheme is quite efficient for several reasons. First, the whole-body landmark detectors are learned offline. Second, whole-body landmark detection only needs to be run once, and then the landmarks can be used for finding corresponding point for any POI clicked in the reference domain. Third, salient region features can be pre-computed in the whole-body, so that those close to a POI can be identified quickly. It should be noted that the simple matching algorithm between salient point features has very low computational complexity and it is meant to provide only rough results. It is thus likely that Q contains non-negligible number of false correspondences. The robust computational steps that follow will allow us to remedy the effect of such possible false correspondences.

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 $\mathbf{c}_f$ as a L-component Gaussian mixture. First we construct a set of all K-subsets of Q. We denote such a set 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. When the cardinality of Q is large, P can be randomly sampled, resulting in $L < \binom{M}{K}$. The L-component Gaussian mixture given $\mathbf{c}_r$ and P is then defined by

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

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

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

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

where f_t and g_t are the estimator of the mean and the width of the i-th Gaussian component given a K-subset P_l that is randomly sampled from Q. Through a geometrical interpretation, we derive closed-form formulae of f_t and g_t by exploiting geometric invariance between a pair of polyhedra under a specific group of domain transformation. The form of f_t and g_t depends on the domain's dimensionality and the type of transformation $\mathcal{T}_\theta$. In this paper, we consider solutions in $\mathbb{R}^3$ for two transformation groups, i) pure translation and ii) scaling and translation, although their extension to more complex projective transformation is also possible using the same strategy.

First suppose that true domain transform $\mathcal{T}_\theta$ is an instance of a specific linear transformation group. Then we can choose the value of K such that the correspondences in a P_l can sufficiently constrain the full degrees of freedom of the transformation, similar to the well-known RANSAC setup [5]. Next 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, these polyhedra must satisfy certain geometric invariances under the supposed transformation group, resulting in a set of equations that must hold true. f_t is then given by solving such equations explicitly about $\mathbf{c}_f$. The following demonstrates this procedure in some details.

Let us arbitrarily pick a vertex pair in P_l and denote it $\mathbf{q}_l = (\mathbf{q}_{rl}, \mathbf{q}_{fl})$. We define a pair of local coordinate frames for both domains by setting their origin at $(\mathbf{q}_{rl}, \mathbf{q}_{fl})$. Then

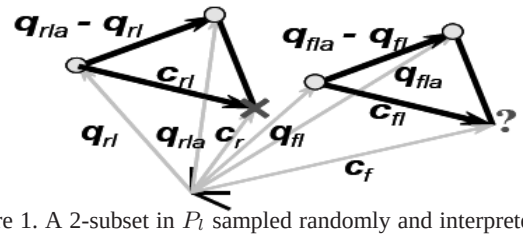

Figure 1. A 2-subset in P_l sampled randomly and interpreted geometrically as a pair of similar triangles. Circles, cross and question mark denote anchor features, click-point and link point, respectively. The triangles are uniquely represented by either three vertices in a global frame, as shown in the light-color vectors, or two vectors in the respective local frames centered at $(\mathbf{q}_{rl}, \mathbf{q}_{fl})$, as shown in the dark-color vectors.

$\mathbf{c}_r$ and $\mathbf{c}_f$ can be described by position vectors $\mathbf{c}_{rl}$ and $\mathbf{c}_{fl}$ in their respective local frame.

$$\mathbf{c}_r = \mathbf{c}_{rl} + \mathbf{q}_{rl}$$

$$\mathbf{c}_f = \mathbf{c}_{fl} + \mathbf{q}_{fl}$$

Figure 1 illustrates these position vectors and local frames. $\mathbf{c}_{fl}$ 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 supposed pure translation, resulting in an equation $\mathbf{c}_{fl} = \mathbf{c}_{rl}$. The solution immediately gives

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

When $K = 2$, each P_l yields two correspondences providing 6 constraints, which are interpreted as a pair of geometrically similar triangles in $\mathbb{R}^3$ as shown in Figure 1. 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. The similar triangles are described by two corresponding vectors $(\mathbf{q}_{rla} - \mathbf{q}_{rl}, \mathbf{c}_{rl})$ and $(\mathbf{q}_{fla} - \mathbf{q}_{fl}, \mathbf{c}_{fl})$. This transformation group assures invariance of corresponding normalized vectors and ratio of corresponding vector norms, resulting in

$$\frac{\mathbf{c}_{fl}}{\|\mathbf{c}_{fl}\|} = \frac{\mathbf{c}_{rl}}{\|\mathbf{c}_{rl}\|}, \quad \frac{\|\mathbf{c}_{fl}\|}{\|\mathbf{q}_{fla} - \mathbf{q}_{fl}\|} = \frac{\|\mathbf{c}_{rl}\|}{\|\mathbf{q}_{rla} - \mathbf{q}_{rl}\|}$$

where $\|\cdot\|$ denote a vector norm. Simple algebra reveals that the desired estimator of the l-th Gaussian component mean with $K = 2$ is derived as follows.

$$\begin{aligned} \mathbf{m}_{l,K=2} &= f_{t,K=2}(\mathbf{c}_r, P_l) \\ &= \frac{\|\mathbf{q}_{fla} - \mathbf{q}_{fl}\|}{\|\mathbf{q}_{rla} - \mathbf{q}_{rl}\|}(\mathbf{c}_r - \mathbf{q}_{rl}) + \mathbf{q}_{fl} \end{aligned} \quad (7)$$

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.

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

Finally, given a sequence of P_l's in P, the successive application of f_t and g_t results in a set of L link estimates forming the Gaussian mixture spatial likelihood. This linking framework can be interpreted as an extension of the well-known RANSAC. While the original RANSAC first estimates a set of domain transformations via random sampling, our solution computes a set of point-wise link estimates using the equivalent sampling scheme but implicitly using the knowledge of domain transformation as constraints. This is a natural extension because the goal of click-point linking is to find a point-wise correspondence rather than the domain transform that is the goal of the RANSAC.

2.3. Mean Shift-based Robust Maximum Likelihood Estimation

This section describes our robust and efficient solution for the maximum likelihood estimation (MLE) problem in (1) with the likelihood model (2-5), providing the final linking estimate. 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 [6]. VBMS is an 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. In comparison the the original RANSAC, this can be interpreted as an application of the mean shift to the consensus step of the RANSAC.

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

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

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 $\mathbf{m}_v(\mathbf{x}) = H_h(\mathbf{x}) \frac{\widehat{\nabla} f_v(\mathbf{x})}{\hat{f}_v(\mathbf{x})}$. Then the following iterative algorithm is provably convergent to a density mode in the vicinity of the initialization $\mathbf{x}_{init}$ in the gradient-ascent sense but without nuisance parameter tuning

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

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

We apply VBMS to our problem by 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 MLE by VBMS:

B1 Compute the means $\mathbf{z}_r$ and $\mathbf{z}_f$ of salient-region feature points in C_r and C_f, respectively.

B2 Compute the mean bias $\mathbf{z} = \mathbf{z}_f - \mathbf{z}_r$ between C_r and C_f.

B3 Set the initialization of a VBMS iterator by the mean bias-corrected POI in the floating domain: $\mathbf{x}_{init} = \mathbf{c}_r + \mathbf{z}$

B4 Perform the VBMS algorithm in (10), resulting in the convergence $\mathbf{y}^*$.

B5 Results in the linking estimate $\widehat{\mathbf{c}}_f = \mathbf{y}^*$.

3. Experimental Studies

We evaluate the proposed framework by testing our 3D implementation 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.

The following setting of the proposed algorithm was used. For each volume, a number of 18 whole-body landmarks are detected. The two volumes are globally aligned

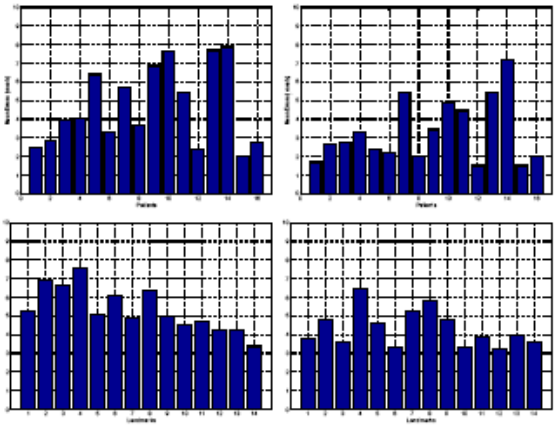

Figure 3. Experimental results. Top: average errors as a function of 16 dierent patients. Bottom: average errors as a function of 14 dierent click points. For feature matching, an unbiased linear combination of the geometric and appearance similarity (2 intensity histogram distance) is used as similarity function. Left: with only salient region features. Right: with global-to-local landmarks. All the errors are calculated with the unit of voxels.

based on whole-body landmark correspondences. A number of 50 salient region features are also pre-computed. For each click-point $\mathbf{c}_r$, besides 5 nearest whole-body landmark correspondences, the salient region feature matching algorithm is performed with 10 nearest salient features: $n = 5, m = 10$. For salient region feature matching, two similarity functions are considered in this study: geometric Euclidean distances and the χ^2 distance of intensity histograms. A solution for scaling and translation with $K = 2$ is considered. For testing, we used pre-recorded 3D point correspondences that are manually labeled by experts. There were 14 points for each volume distributed in 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 links in the floating domain of $\mathbb{R}^3$. The total of 224 test cases (16 patients over 14 points) 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 point was automatically estimated by using the maximum entropy criterion [4].

Figure 2 shows some illustrative examples. For the shoulder case, significant dissimilarity of central torso region indicates a body twist between the image pair. This exemplifies the advantage of the click-point linking approach over the rigid global registration which cannot assure specific locations to be accurate. In the follow-up settings, tumor and heart are regions that often change in appearance as are shown in the figure. Our results show successful linking, indicating the effectiveness of this approach.

Fig. 3 shows the result of our experiments. The top

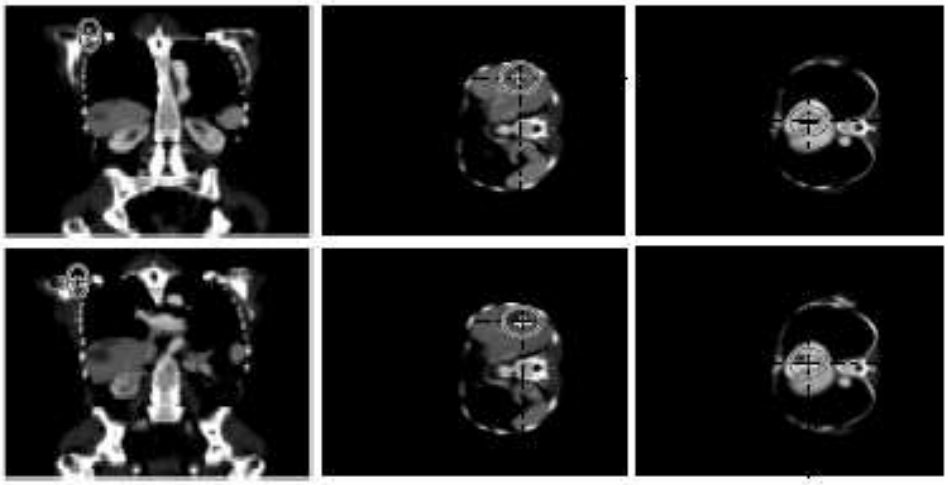

Figure 2. Three illustrative examples. Top: reference images with POI. Bottom: floating images with link estimates. Ellipses indicate saliency feature scales shown with anisotropic voxels. Left: shoulder bones. Middle: liver tumor. Right: heart.

row shows the average errors plotted over different patients. The bottom row shows those plotted over different click points. For feature correspondence matching, we consider a similarity function as a linear combination of geometric Euclidean distance with the mean bias adjustment and an appearance-based distance using χ^2 distance of intensity histograms. We compare the results of our system with another system using only salient region features shown in [10]. The left column shows the results with the system using 10 salient region features only. The total average and median errors were 4.68 and 3.10 voxels, respectively. On the other hand, the results with the proposed system using the global-to-local landmark hierarchy are shown in the right column. The average and median errors were 3.78 and 2.70, respectively. For extracting 18 whole-body landmarks and 50 salient region features in a 3D volume with 128 by 128 slices, it took roughly 2.8 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 click points. Visual inspec- tion revealed that higher errors (e.g, patient 7 and 14) were caused mainly by the outlier failures due to lack of corresponding salient region 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 that extends the RANSAC to our linking problem. Given a set of corresponding features found by the cascade and entropy-based detectors, the geometrical context of arbitrary POI is modeled by a Gaussian mixture spatial likelihood built by using a RANSAC-type random sampling. Then variable bandwidth mean shift is utilized for solving the MLE problem robustly and efficiently. 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 feature extraction and matching parts.

References

[1] Brown, L.G.: A survey of image registration techniques. ACM Comput. Surv. **24**(4) (1992) 325–376 2

[2] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. (2001) 511–518 2, 3

[3] Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision **45**(2) (2001) 83–105 2, 4

[4] Huang, X., Sun, Y., Metaxas, D., Sauer, F., Xu, C.: Hybrid image registration based on configural matching of scale-invariant salient region features. In: IEEE Workshop on Image and Video Registration. (2004) 2, 4, 6

[5] 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 2, 4

[6] Comaniciu, D.: An algorithm for data-driven bandwidth selection. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(2) (2003) 281–288 2, 5

[7] Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. Volume 2. (2003) 264–271 2

[8] Epshtein, B., Ullman, S.: Identifying semantically equivalent object fragments. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. Volume 1. (2005) 2–9 2

[9] Novak, C., Shen, H., Odry, B., Ko, J., Naidich, D.: System for automatic detection of lung nodules exhibiting growth. In: SPIE Med. Imag. (2004) 2

[10] Okada, K., Huang, X., Zhou, X., Krishnan, A.: Robust click-point linking for longitudinal follow-up studies. In: Proc. of Intl Workshop on Medical Imaging and Augmented Reality. (2006) 2, 4, 7

[11] Frangi, A.F., Rueckert, D., Schnabel, J.A., Niessen, W.J.: Automatic 3d asm construction via atlas-based landmarking and volumetric elastic registration. In: Proc. of Information Processing in Medical Imaging. (2001) 78–91 2

[12] Rueckert, D., Frangi, A., Schnabel, J.: Automatic construction of 3d statistical deformation models using non-rigid registration. In: Proc. of International Conf. on Medical Imaging Copmuting and Computer-Assisted Intervention. (2001) 77–84 2

[13] Chen, M., Kanade, T., Pomerleau, D., Schneider, J.: 3-D deformable registration of medical images using a statistical atlas. In: Proc. of International Conf. on Medical Imaging Copmuting and Computer-Assisted Intervention. (1999) 621–630 2

[14] Thirion, J.P.: New feature points based on geometric invariants for 3D image registration. International Journal of Computer Vision **18**(2) (1996) 121–137 3

[15] Can, A., Stewart, C., Roysam, B., Tanenbaum, H.: A feature-based, robust, hierarchical algorithm for registering pairs of images of the curved human retina. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(3) (2002) 347–364 3

[16] Maurer, C., Maciunas, R., Fitzpatrick, J.: Registration of head CT images to physical space using a weighted combination of points and surfaces. IEEE Transactions on Medical Imaging **17**(5) (1998) 753–761 3

[17] 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 3

[18] Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models - their training and application. Computer Vision and Image Understanding **61**(1) (1995) 38–59 3

[19] Fitzgibbon, A.: Robust registration of 2D and 3D point sets. In: Proc. of British Machine Vision Conference. Volume 2. (2001) 411–420 4