Task-Optimal Registration Cost Functions

B.T. Thomas Yeo[1,*], Mert Sabuncu[1], Polina Golland[1], and Bruce Fischl[2,1]

[1] Computer Science and Artificial Intelligence Laboratory, MIT, USA
[2] Athinoula A. Martinos Center for Biomedical Imaging, MGH/HMS, USA
ythomas@csail.mit.edu

Abstract. In this paper, we propose a framework for learning the parameters of registration cost functions – such as the tradeoff between the regularization and image similiarity term – with respect to a specific task. Assuming the existence of labeled training data, we specialize the framework for the task of localizing hidden labels via image registration. We learn the parameters of the weighted sum of squared differences (wSSD) image similarity term that are optimal for the localization of Brodmann areas (BAs) in a new subject based on cortical geometry. We demonstrate state-of-the-art localization of V1, V2, BA44 and BA45.

1 Introduction

In medical imaging, registration is rarely the end-goal, and therefore the quality of image registration should be evaluated in the context of the application. The results of registration are usually used by other tasks, e.g., segmentation. Taking into account the parameters of the registration cost function has been shown to improve alignment as measured by the performance of subsequent tasks, such as population analysis [1] and segmentation [2,3]. This paper proposes a framework for optimizing parameters of registration cost functions for a specific task.

A common image similarity measure used in registration is the weighted sum of squared differences (wSSD). wSSD assumes an independent Gaussian distribution on the image noise with the weights corresponding to the reciprocal of the variance. The weights are typically set to a constant global value [4,5]. Alternatively, a spatially-varying variance can be estimated from the intensities of registered images [6]. However, the estimated variance depends on the wSSD-regularization tradeoff: weaker deformation regularization leads to better intensity alignment and lower variance estimates [3].

Recent work in probabilistic template construction resolves this problem by marginalizing the tradeoff under a Bayesian framework [7] or estimating the tradeoff with the Minimum Description Length principle [8]. Since these methods are not guided by any task, it is unclear whether the resulting parameters are optimal for any specific task. After all, the optimal parameters for segmentation might be different from those for group analysis. In contrast, [9] proposes a generative model for segmentation, so the registration parameters are Bayesian-optimal for segmentation. When considering a single global tradeoff parameter,

---

[*] Corresponding author.

an exhaustive search by cross-validation of segmentation accuracy is possible [3]. Unfortunately, an exhaustive search is not feasible for multiple parameters.

Unlike these generative approaches, we take the discriminative approach of directly incorporating the task in the parameter estimation. We assume that the performance of a particular task can be measured by a cost function g given the output of registration. Our method learns the parameters of a given registration cost function f that yield better registration of a new image with respect to a specific task. The task-specific cost function g is evaluated with information from training data that is not available to the registration cost function f.

Our formulation is related to the computation of the entire space of solutions of learning problems (e.g. SVM) as a function of a single regularization parameter [10]. Because we deal with multiple parameters, it is impossible for us to compute a solution manifold. Instead, we trace a path within the solution manifold that improves the task-specific cost function.

In this paper, we propose a framework that optimizes parameters of registration cost functions for a specific task. We learn the weights of the wSSD registration cost function to optimize the prediction of Brodmann Areas (BAs) in a new subject, effectively estimating the tradeoff between the similarity measure and regularization. We demonstrate improvement over existing localization methods [11] for several BAs.

2   Task-Optimal Registration

Given an atlas coordinate frame and a new image, $f(w, \Psi)$ denotes a smooth registration cost function parametrized by the weights w and transformation Ψ. For example, the parameters w can be the tradeoff between the regularization and image similarity measure. f is typically a function of the template and the input image, but we omit this dependency to simplify notation. We assume a known and fixed template. Image registration is the process of estimating Ψ^* for a given set of parameters w:

$$\Psi^*(w) = \operatorname*{argmin}_{\Psi} f(w, \Psi). \tag{1}$$

A different set of parameters w will result in a different solution and thus will effectively lead to a different image coordinate system. While there are typically multiple solutions to Eq. (1), we work with a single local optimum in practice.

The results of registration are used for further tasks, such as image segmentation. We assume the task performance can be measured by a smooth cost function g, so that smaller values of $g(\Psi^*(w))$ correspond to better task performance. g is a function of input data associated with a subject, such as its anatomical labels or functional activation map, not available to the cost function f.

Given a set of training subjects, we seek the parameters w^* that generalize well to a new subject, i.e., registration of a new subject with w^* yields the transformation $\Psi^*(w^*)$ with a small task-specific cost $g(\Psi^*(w^*))$:

$$w^* = \operatorname*{argmin}_{w} G(w) \qquad \text{where} \qquad G(w) \triangleq \sum_{s=1}^{S} g_s(\Psi_s^*(w)) + \operatorname{Reg}(w). \tag{2}$$

$\Psi_s^*(w)$ denotes the registration of training subject s with fixed weights w. g_s is the evaluation of g for training subject s. $\text{Reg}(w)$ denotes regularization on w.

2.1   Optimizing Registration Parameters w

In this section, we discuss the optimization of the cost function in Eq. (2). Let ∂_x, ∂_x^2, and $\partial_{x,y}^2$ denote partial derivatives and $\Psi^*(w_0)$ denote a local minimum of the registration cost function for a fixed $w = w_0$.

Suppose we perturb w_0 by δw. Let $\Psi^*(w_0) + \delta\Psi^*(w_0, \delta w)$ denote the new locally optimal deformation for the new parameters $w_0 + \delta w$. Because of numerous local optima, $\Psi^*(w_0) + \delta\Psi^*(w_0, \delta w)$ might be far from $\Psi^*(w_0)$. If the Hessian $\partial_\Psi^2 f(w_0, \Psi)$ is positive definite at $\Psi = \Psi^*(w_0)$, then by the Implicit Function Theorem [12], a unique function $\delta\Psi^*(w_0, \delta w)$ exists with the same order of smoothness as f for small enough values of $\|\delta w\|$, such that $\delta\Psi^*(w_0, 0) = 0$.

Consequently, at $(w_0, \Psi^*(w_0))$ with positive definite Hessian, one can compute the derivatives of $\delta\Psi^*$, allowing us to traverse a curve of local optima, finding values of w that improve the task-specific cost function for the training images. Since the derivatives at any local optimum is zero, we can show that

$$\partial_w \Psi^* \Big|_{w_0} = - \left(\partial_\Psi^2 f(w_0, \Psi) \Big|_{\Psi^*(w_0)} \right)^{-1} \partial_{w,\Psi}^2 f(w, \Psi) \Big|_{w_0, \Psi^*(w_0)}. \tag{3}$$

In practice, the matrix inversion in Eq. (3) is computationally prohibitive for high-dimensional warps Ψ. As a result, we consider a simplification of Eq. (3) by setting the Hessian to be the identity:

$$\partial_w \Psi^* \Big|_{w_0} \approx - \partial_{w,\Psi}^2 f(w, \Psi) \Big|_{w_0, \Psi^*(w_0)}. \tag{4}$$

Since $-\partial_\Psi f$ is the direction of gradient descent of the cost function Eq. (1), we can interpret Eq. (4) as approximating the new local minimum to be in the same direction as the *change* in the direction of gradient descent as w is perturbed. Differentiating the cost function in Eq. (2), using the chain rule, we get

$$\partial_w G = \partial_w \left(\sum_{s=1}^{S} g_s(\Psi_s^*(w)) + \text{Reg}(w) \right) = \sum_{s=1}^{S} [\partial_{\Psi_s^*} g_s][\partial_w \Psi_s^*] + \partial_w \text{Reg}(w). \tag{5}$$

We can therefore optimize Eq. (2) by standard gradient descent. We summarize the training procedure of the task-optimal image registration framework below:

- Initialize w to uniform values.
- Estimate $\Psi_s^*(w) = \text{argmin}_{\Psi_s} f_s(w, \Psi_s)$, i.e., perform registration of each training subject s.
- Iterate until convergence:
  - Given current estimates $(w, \{\Psi_s^*(w)\})$, compute the gradient $\partial_w G$ in Eq. (5) using $\partial_w \Psi^*$ in Eq. (4).
  - Perform line search in the direction opposite to the gradient $\partial_w G$.

Each line search involves evaluating the cost function G multiple times, which in turn requires registering the training subjects. Since we are initializing from a local optimum, for a small change in w, the registration converges quickly.

Since nonlinear registration is dependent on initialization, the current estimates $(w, \Psi^*(w))$, which were initialized from previous estimates, might not be achievable when initializing the registration with the identity transform. The corresponding parameters w might therefore *not* generalize well to a new subject initialized with the identity transform. Consequently, after every few iterations, we re-initialize the transformations to the identity transform, re-register the images and check that G is better than the current best value of G computed with initialization from the identity transform.

Remark: Degeneracies can arise for local minima with a singular Hessian. For example, let $\Psi = [a \ b]$ and $f(\Psi, w) = (ab - w)^2$. Then the determinant of the Hessian at *any* local minimum is equal to zero! In this case, there is an infinite number of local minima near the current local minimum $\Psi^*(w)$, i.e., the gradient is not defined - our algorithm might then be stuck in the current estimates of w. In our experiments, these degeneracies do not seem to pose serious problems.

3   Learning wSSD for Hidden Label Alignment

We now instantiate the task-optimal registration framework for localizing hidden labels in images. Here, we work with meshes modeling the cortical surface, although it should be clear that the discussion extends to volumetric images. We assume that the meshes have been spherically parameterized and represented as spherical images: a geometric attribute is associated with each mesh vertex, describing local cortical geometry.

Suppose we have a set of spherical training images $\{I_s\}$ with a particular structure manually labeled. We represent the binary labels as signed distance transforms on the sphere $\{L_s\}$. We assume the existence of a spherical image template I_T and corresponding distance transform L_T. In this paper, we select one of the training subjects as the template. Our task is to align a new image to the template and predict the boundary of the hidden structure in the new subject by transferring the labels from the template to the new subject.

We represent the transformation Ψ as a composition of diffeomorphic warps, each parameterized by a stationary velocity field [5,13]. A diffeomorphic warp Φ is associated with a smooth stationary velocity field v via a stationary ODE: $\partial_t \Phi(x, t) = v(\Phi(x, t))$ with an initial condition $\Phi(x, 0) = x$. The solution at $t = 1$ is denoted as $\Phi(x, 1) = \Phi(x) = \exp(v)(x)$, where we have dropped the time index. A solution can be computed efficiently using scaling and squaring [14]. For a given image I_s, we define the registration cost function:

$$f_s(w, \Psi) = \sum_i w_i^2 \left[I(x_i) - I_s(\Psi(x_i)) \right]^2 + \sum_i \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \left(\frac{\|\Psi(x_i) - \Psi(x_j)\| - d_{ij}}{d_{ij}} \right)^2,$$

where $\Psi(x_i)$ denotes the point on the sphere S^2 to which Ψ maps the point $x_i \in S^2$. The first term corresponds to the wSSD image similarity. The

parameterization of the weights as w_i^2 ensures non-negative weights. The second term is the regularization on the transformation Ψ. $\mathcal{N}_i$ is a predefined neighborhood around vertex i, d_{ij} is the original distance between the neighbors $\|x_i - x_j\|$. A higher weight w_i corresponds to placing more emphasis on matching the cortical geometry of the template at spatial location x_i relative to the regularization.

To register subject s to the template, let Ψ_0 be the current estimate of Ψ. We seek an update $\exp(v)$ parameterized by a stationary velocity field v:

$$f_s(w, \Psi_0 \circ \exp(v)) = \sum_i w_i^2 \left[I_T(x_i) - I_s(\Psi_0 \circ \exp(v)(x_i)) \right]^2 \tag{6}$$

$$+ \sum_i \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \left(\frac{\|\Psi_0 \circ \exp(v)(x_i) - \Psi_0 \circ \exp(v)(x_j)\| - d_{ij}}{d_{ij}} \right)^2.$$

We adopt the techniques in the Spherical Demons algorithm [13] to differentiate Eq. (6) with respect to v, evaluated at $v = 0$. Defining $\nabla I_s(\Psi_0(x_i))$ to be the gradient of the warped image $I_s(\Psi_0(\cdot))$ at x_i, $\nabla\Psi_0(x_i)$ to be the Jacobian matrix of Ψ_0 at x_i and v_i to be the velocity vector tangent to vertex x_i, we get

$$\partial_{v_i} f_s(w, \Psi_0 \circ \exp(v)) \Big|_{v=0} = -2w_i^2 \left[I_T(x_i) - I_s(\Psi_0(x_i)) \right] \left[\nabla I_s(\Psi_0(x_i)) \right]^T \tag{7}$$

$$+ 2 \sum_{j \in \mathcal{N}_i} \left(\frac{1}{|\mathcal{N}_i|} + \frac{1}{|\mathcal{N}_j|} \right) \left(\frac{\|\Psi_0(x_i) - \Psi_0(x_j)\| - d_{ij}}{d_{ij}^2 \|\Psi_0(x_i) - \Psi_0(x_j)\|} \right) \left[\Psi_0(x_i) - \Psi_0(x_j) \right]^T \nabla\Psi_0(x_i).$$

Eq. (7) instantiates $\partial_\Psi f_s$ for this application. We can then perform gradient descent of the registration cost function f_s to obtain Ψ_s^*, which can be used to evaluate the task-specific cost function g_s. We adopt a simple label similarity measure for our task of localizing hidden labels:

$$g_s(\Psi^*) = \sum_i \left[L_T(x_i) - L_s(\Psi_s^*(x_i)) \right]^2, \tag{8}$$

A low value of g_s indicates good alignment of the hidden label maps between the template and subject s, suggesting good prediction of the hidden label.

Here, we ignore the regularization $\text{Reg}(w)$, but still achieve good results. One reason is that the re-registration after every few line searches helps to regularize against bad values of w. There is also implicit regularization in the framework: for example, w cannot become arbitrary large, since registration achieved with almost no regularization will lead to poor task performance.

Given the current estimates (w, Ψ_s^*), to update w using Eq. (5), we evaluate:

$$\partial_{\Psi^*} g_s = \partial_{v_i} g_s(\Psi_s^* \circ \exp(v)) \Big|_{v=0} = -2 \left[L_T(x_i) - L_s(\Psi_s^*(x_i)) \right] \left[\nabla L_s(\Psi_s^*(x_i)) \right]^T \tag{9}$$

$$\partial_w \Psi_s^* \approx -\partial_{w_i, v_j}^2 f_s(w, \Psi_s^* \circ \exp(v)) \Big|_{v=0} = 4w_i \left[I(x_i) - I_s(\Psi_s^*(x_i)) \right] \left[\nabla I_s(\Psi_s^*(x_i)) \right]^T \delta(i, j).$$

4   Experiments

In this section, we demonstrate the utility of the task-optimal registration framework for localizing Brodmann Areas (BAs). We compare the framework with using uniform weights [4,5] and FreeSurfer [6].

BAs are cyto-architectonically defined parcellations of the cerebral cortex closely related to cortical function. We consider 10 human brains analyzed via postmortem histology [15]. Histologically defined BAs were sampled onto each hemispheric surface model and sampling errors were manually corrected. In this paper, we consider V1 and V2, which are well-predicted by local geometry and the Broca's areas: BA44 and BA45, which are not [11].

Even though each subject has multiple BAs, we focus on each structure independently. This allows us to interpret the weights in the wSSD in association with a particular label: a large weight at a particular location implies that the cortical geometry at that spatial location of the template is significant for localizing the label of interest.

4.1   Methods

Task-Optimal. We perform leave-one-out cross validation to predict BA location. For each BA and a test subject, we use one of the remaining 9 subjects as the template and the remaining 8 subjects for training. Once the weights are learned, we use them to register the test subject and predict the BA of the test subject by transferring the BA label from the template to the subject. We compute the symmetric mean Hausdorff distance between the boundary of the true BA and the predicted BA on the cortical surface of the test subject – smaller Hausdorff distance corresponds to better localization. There are 90 possibilities to select the test subject and the template. Here, we consider 20 of the 90 possibilities by selecting each of the 10 subjects to be a test subject twice (with a different randomly selected template), resulting in a total of 20 trials and 20 mean Hausdorff distances for each BA and for each hemisphere.

Uniform-weight. We repeat the process for the uniform-weight method using the same 20 pairs of subjects, where all the w_i's are manually set to a global fixed value w without training. We explore 12 different values of global weight w, chosen so that the deformations range from rigid to flexible warps. For each BA and each hemisphere, we pick the *best* value of w leading to the lowest mean Hausdorff distances. Because there is no cross-validation in picking the weights, the uniform-weight method is using an unrealistic version of the strategy proposed in [3].

FreeSurfer. Finally, we use FreeSurfer [6] to register the 10 *ex vivo* subjects to the FreeSurfer Buckner40 atlas, constructed from the MRI of 40 *in vivo* subjects. Once registered into this *in vivo* atlas space, for the same 20 pairs of subjects, we can use the BAs of one *ex vivo* subject to predict another *ex vivo* subject. We note that FreeSurfer also uses the wSSD cost function, but a more sophisticated regularization that penalizes both metric and areal distortion. For a particular tradeoff between the similarity measure and regularization, the Buckner40 template consists of the empirical mean and variance of the 40 *in vivo* subjects registered to template space. We use the reported FreeSurfer tradeoff parameters that were used to produce prior state-of-the-art BA alignment [11].
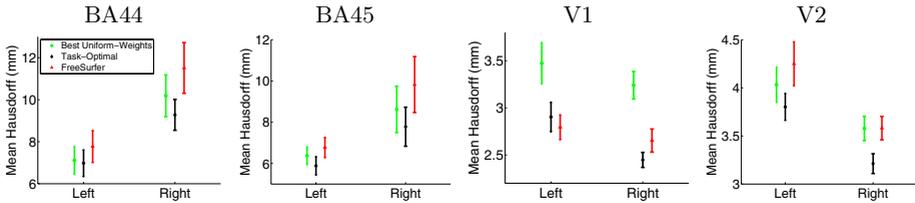
Fig. 1. Mean Hausdorff Distance (in mm) for V1, V2, BA44 and BA45. For the uniform-weight method, the result corresponding to the *best* weight is reported.

We run the task-optimal and uniform-weight methods on a low-resolution subdivided icosahedron mesh containing 2,562 vertices, whereas FreeSurfer results were computed on high-resolution meshes of more than 100k vertices. In our implementation, training on 8 subjects takes on average 4hrs on a standard PC (AMD Opteron, 2GHz, 4GB RAM). Despite the use of the low-resolution mesh, we achieve state-of-the-art localization accuracy.

4.2   Results

Fig. 1 shows the alignment errors for V1, V2, BA44 and BA45. Not surprisingly, we achieve better localization of BA44 and BA45 over the uniform-weight method and FreeSurfer, since local geometry poorly predicts the Broca's areas.

Since local cortical geometry is predictive of V1 and V2, we expect the three methods to perform similarly for V1 and V2. Surprisingly, we achieve improvement in V2 alignment over the uniform-weight method and FreeSurfer. Our method also significantly improves the alignment of V1 with respect to the uniform-weight method. Compared with FreeSurfer, we achieve slightly worse localization in the left hemisphere but better localization in the right. This suggests that even when local geometry is predictive of the hidden labels, so that anatomy-based registration is reasonable for localizing the labels, tuning the registration cost function can further improve the task performance.

Since our measure of localization accuracy uses the mean Hausdorff distance, ideally we should incorporate it into our task-specific objective function instead of the SSD on the distance transform representing the BA. Unfortunately, the resulting derivative is difficult to compute and the gradient will be zero everywhere except at the BA boundaries, resulting in a slow optimization.

We note that the task-optimal and uniform-weights registrations are pairwise, while FreeSurfer registrations are performed via an atlas. In our experience, registration via an unbiased atlas is usually more accurate than direct pairwise registration. Furthermore, FreeSurfer utilizes atlas-based registration by default, and this was used to produce prior best BA alignment [11]. We also note that our approach allows the computation of multiple task-optimal templates, thus complementing recent approaches of using multiple atlases for segmentation [16,17].

5   Conclusion

In this paper, we present a framework for optimizing the parameters of any differentiable family of registration cost functions with respect to a specific task. The only requirement is that the task performance can be encoded by a differentiable cost function. We demonstrate state-of-the-art Brodmann area localization by optimizing the weights of the wSSD image-similarity measure. Future work involves applying the framework to other Brodmann areas and fMRI-defined functional regions, as well as estimating the optimal template in addition to the weights of the registration cost function. We also hope to design task-specific cost functions for tasks other than segmentation.

References

1. Makrogiannis, S., Verma, R., Karacali, B., Davatzikos, C.: A joint transformation and residual image descriptor for morphometric image analysis using an equivalence class formulation. In: Proc. MMBIA (2006)
2. Commowick, O., Stefanescu, R., Fillard, P., Arsigny, V., Ayache, N., Pennec, X., Malandain, G.: Incorporating statistical measures of anatomical variability in atlas-to-subject registration for conformal brain radiotherapy. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3750, pp. 927–934. Springer, Heidelberg (2005)
3. Yeo, B., Sabuncu, M., Desikan, R., Fischl, B., Golland, P.: Effects of Registration Regularization and Atlas Sharpness on Segmentation Accuracy. Medical Image Analysis 12(5), 603–615 (2008)
4. Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased Diffeomorphic Atlas Construction for Computational Anatomy. NeuroImage 23, 151–160 (2004)
5. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: Efficient non-parametric image registration. NeuroImage 45(1), S61–S72 (2009)
6. Fischl, B., Sereno, M., Tootell, R., Dale, A.: High-resolution Intersubject Averaging and a Coordinate System for the Cortical Surface. Human Brain Mapping 8(4), 272–284 (1999)
7. Allassonniere, S., Amit, Y., Trouvé, A.: Toward a Coherent Statistical Framework for Dense Deformable Template Estimation. Journal of the Royal Statistical Society, Series B 69(1), 3–29 (2007)
8. Twining, C., Cootes, T., Marsland, S., Petrovic, V., Schestowitz, R., Taylor, C.: A Unified Information-Theoretic Approach to Groupwise Non-rigid Registration and Model Building. In: Christensen, G.E., Sonka, M. (eds.) IPMI 2005. LNCS, vol. 3565, pp. 1611–3349. Springer, Heidelberg (2005)

9. Van Leemput, K.: Probabilistic Brain Atlas Encoding Using Bayesian Inference. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 704–711. Springer, Heidelberg (2006)
10. Park, M., Hastie, T.: l_1-regularization Path Algorithm for Generalized Linear Models. J. of the Royal Stat. Soc. B 69, 659–677 (2007)
11. Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B., Mohlberg, H., Amunts, K., Zilles, K.: Cortical Folding Patterns and Predicting Cytoarchictecture. Cerebral Cortex 18(8), 1973–1980 (2008)
12. Rudin, W.: Principles of Mathematical Analysis (1976)
13. Yeo, B., Sabuncu, M., Vercauteren, T., Ayache, N., Fischl, B., Golland, P.: Spherical Demons: Fast Surface Registration. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 745–753. Springer, Heidelberg (2008)
14. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A Log-Euclidean Framework for Statistics on Diffeomorphisms. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 924–931. Springer, Heidelberg (2006)
15. Amunts, K., Schleicher, A., Burgel, U., Mohlberg, H., Uylings, H., Zilles, K.: Broca's Region Revisited: Cytoarchitecture and Intersubject Variability. Journal of Comparative Neurology 412(2), 319–341 (1999)
16. Heckemann, R., Hajnal, J., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain mri segmentation combining label propagation and decision fusion. NeuroImage 33(1), 115–126 (2006)
17. Sabuncu, M., Balci, S., Golland, P.: Discovering Modes of an Image Population through Mixture Modeling. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part II. LNCS, vol. 5242, pp. 381–389. Springer, Heidelberg (2008)