Determining Functional Units of Tongue Motion via Graph-Regularized Sparse Non-negative Matrix Factorization

Jonghye Woo[1,2], Fangxu Xing[2], Junghoon Lee[2], Maureen Stone[1], and Jerry L. Prince[2]

[1] University of Maryland, Baltimore MD, USA
[2] Johns Hopkins University, Baltimore MD, USA
jschant@gmail.com

Abstract. Tongue motion during speech and swallowing involves synergies of locally deforming regions, or functional units. Motion clustering during tongue motion can be used to reveal the tongue's intrinsic functional organization. A novel matrix factorization and clustering method for tissues tracked using tagged magnetic resonance imaging (tMRI) is presented. Functional units are estimated using a graph-regularized sparse non-negative matrix factorization framework, learning latent building blocks and the corresponding weighting map from motion features derived from tissue displacements. Spectral clustering using the weighting map is then performed to determine the coherent regions—i.e., functional units—defined by the tongue motion. Two-dimensional image data is used to verify that the proposed algorithm clusters the different types of images accurately. Three-dimensional tMRI data from five subjects carrying out simple non-speech/speech tasks are analyzed to show how the proposed approach defines a subject/task-specific functional parcellation of the tongue in localized regions.

1 Introduction

The relationship between the structural and functional components of the tongue is poorly understood partly due to the complex tongue anatomy and muscle interactions. The human tongue is a volume preserving structure with highly complex, orthogonally oriented, and interdigitated muscles. The tongue muscles interact with one another in order to carry out the oromotor behaviors of speaking, swallowing, and breathing, which are executed by deforming local functional units in this complex muscular array. Tongue motions are synergies created by locally deforming regions, or *functional units* [1]. Functional units are regions of the tongue that exhibit homogeneous motion during the execution of the specific task. Therefore, identifying functional units and understanding the mechanisms of coupling among them can identify motor control strategy in both normal and adapted speech (e.g., tongue motion after tongue cancer surgery).

To understand the function of the tongue, magnetic resonance imaging (MRI) has played a pivotal role in imaging both tongue surface motion using cine-MRI [2, 3] and internal tissue motion using tagged-MRI (tMRI) [4]. Despite

P. Golland et al. (Eds.): MICCAI 2014, Part II, LNCS 8674, pp. 146–153, 2014.

the rich data on internal tissue motion that is available from tMRI, there has been very little research on its analysis to determine functional units. A key previous report is that of Stone et al. [5] who presented a method to determine functional segments using ultrasound and tMRI. Another key report is that of Ramanarayanan et al. [6], who used a convolutive non-negative matrix factorization (NMF) algorithm to determine tongue movement primitives from electromagnetic articulatory. Our work is inspired by both approaches, but we use richer tMRI-derived data (3D displacements) and the NMF approach with the addition of sparsity and intrinsic data geometry in defining functional units.

Modeling a data matrix as sparse linear combinations of basis vectors is a popular approach to understanding speech production. Among them, NMF and variants involving sparsity have received substantial attention since the seminal work by Lee and Seung [7]. NMF is a matrix factorization method that focuses on data matrices whose elements are non-negative. NMF is based on a parts-based representation inspired by psychological and physiological observations of the brain [8]. However, since standard NMF assumes a standard Euclidean distance measure for its data, it fails to discover the intrinsic geometry of its data [8].

In our work, we assume a manifold of the data within an NMF approach, which thereby captures the intrinsic geometry of the motion features derived from tMRI. In particular, we propose a new approach to determine functional units of tongue motion from tMRI using graph-regularized sparse NMF with spectral clustering. The method integrates a regularization term that encourages the computation of distances on a manifold rather than the whole of Euclidean space in order to preserve the intrinsic geometry of the observed motion data. The use of NMF is important because it does not allow negative combinations of basis vectors. This is consistent with the analysis of muscles, which either have positive activation or no activation, not negative activation. Both quantitative and qualitative evaluation results demonstrate the validity of the proposed method and its superiority to conventional clustering algorithms.

2 Proposed Approach

2.1 Problem Statement

Consider a set of P internal tongue tissue points each with n scalar quantities (e.g., magnitude and angle of each track) tracked through F time frames. These quantities characterize each point and are used to group them into functional units. The location of the p-th tissue point at the f-th time frame can be written as $(x_f^p,\, y_f^p,\, z_f^p)$. The tongue motion can then be represented by a $3F \times P$ spatio-temporal feature matrix $\mathbf{N} = [\mathbf{n}_1, ..., \mathbf{n}_P] \in \mathbb{R}^{3F \times P}$, where the p-th column is given by

$$\mathbf{n}_p = [x_1^p, \cdots, x_F^p, y_1^p, \cdots, y_F^p, z_1^p, \cdots, z_F^p]^T. \tag{1}$$

We cast the problem of determining the functional units as a motion clustering problem. Thus, the goal is to determine a permutation of the columns to form $[\mathbf{N}_1|\, \mathbf{N}_2| \cdots |\mathbf{N}_c]$, where the submatrix $\mathbf{N}_i$ comprises point tracks associated

Algorithm 1. Determination of the functional units

```
1. Extract motion features from displacement fields and construct $\mathbf{U}$.
2. Apply graph-regularized sparse NMF to $\mathbf{U}$ to obtain $\mathbf{V}$ and $\mathbf{W}$.
3. Compute affinity matrix $\mathbf{A}$ from $\mathbf{W}$.
4. Apply spectral clustering to $\mathbf{A}$ and determine functional units.
```

with the i-th submotion—i.e., the i-th functional unit. We provide the proposed approach in more detail. The overall algorithm is shown below.

2.2 Extraction of Motion Quantities

The first step in our algorithm is to extract the motion features that characterize the cohesive motion patterns over time. We extract motion features including the magnitude and angle of the track as in [9] described as

$$m_f^p = \sqrt{(x_{f+1}^p - x_f^p)^2 + (y_{f+1}^p - y_f^p)^2 + (z_{f+1}^p - z_f^p)^2} \tag{2}$$

$$cz_f^p = \frac{x_{f+1}^p - x_f^p}{\sqrt{(x_{f+1}^p - x_f^p)^2 + (y_{f+1}^p - y_f^p)^2}} + 1 \tag{3}$$

$$cx_f^p = \frac{y_{f+1}^p - y_f^p}{\sqrt{(y_{f+1}^p - y_f^p)^2 + (z_{f+1}^p - z_f^p)^2}} + 1 \tag{4}$$

$$cy_f^p = \frac{z_{f+1}^p - z_f^p}{\sqrt{(z_{f+1}^p - z_f^p)^2 + (x_{f+1}^p - x_f^p)^2}} + 1, \tag{5}$$

where m_f^p denotes the magnitude of the track and cz_f^p, cx_f^p, and cy_f^p represent the cosine of the angle projected in the z, x, and y axes plus one, respectively, which are in the range of 0 to 2. For clustering, we gather all the motion features into a $4(F-1) \times P$ non-negative matrix $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_n] \in \mathbb{R}_+^{m \times n}$, where the p-th column can be expressed as

$$\mathbf{u}_p = [m_1^p, \cdots, m_{F-1}^p, cz_1^p, \cdots, cz_{F-1}^p, cx_1^p, \cdots, cx_{F-1}^p, cy_1^p, \cdots, cy_{F-1}^p]^T. \tag{6}$$

These features are always non-negative and can therefore be input to NMF.

2.3 Graph-Regularized Sparse Non-negative Matrix Factorization

NMF: Given a non-negative data matrix $\mathbf{U}$ and $k \leq \min(\mathrm{m}, \mathrm{n})$, let $\mathbf{V} = [v_{ik}] \in \mathbb{R}_+^{m \times k}$ be the *building blocks* and let $\mathbf{W} = [w_{kj}] \in \mathbb{R}_+^{k \times n}$ be the *weighting map.* The goal of NMF is to learn building blocks and corresponding weights such that the input $\mathbf{U}$ is approximated by a product of two non-negative matrices (i.e., $\mathbf{U} \approx \mathbf{VW}$). A typical way to define NMF is to use the Frobenius norm to measure the difference between $\mathbf{U}$ and $\mathbf{VW}$ [7] given by

$$\mathcal{E}(\mathbf{V},\mathbf{W}) = \|\mathbf{U}-\mathbf{V}\mathbf{W}\|_F^2 = \sum_{i,j}\left(u_{ij} - \sum_{k=1}^{K} v_{ik}w_{kj}\right)^2 \tag{7}$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm. The solution can be found through the multiplicative update rule [7]:

$$\mathbf{V} \leftarrow \mathbf{V}.*\mathbf{U}\mathbf{W}^T./\mathbf{V}\mathbf{W}\mathbf{W}^T \tag{8}$$

$$\mathbf{W} \leftarrow \mathbf{W}.*\mathbf{V}^T\mathbf{U}./\mathbf{V}^T\mathbf{V}\mathbf{W} \tag{9}$$

Sparsity Constraint: In this work, we impose a sparsity constraint on the weighting map $\mathbf{W}$. The sparsity constraint allows to encode the high-dimensional motion data using only a few active components, thereby making the weighting map easy to interpret. In particular, the weighting map obtained this way will represent the simplest tongue behavior that could generate the observed motion. In the NMF framework, it has been reported that a fractional regularizer using the $L_{1/2}$ norm outperformed the L_1 norm regularizer and gave sparser solutions [10]. Thus, we incorporate the $L_{1/2}$ sparsity constraint into the NMF framework, which can be expressed as

$$\mathcal{E}(\mathbf{V},\mathbf{W}) = \frac{1}{2}\|\mathbf{U}-\mathbf{V}\mathbf{W}\|_F^2 + \eta\|\mathbf{W}\|_{1/2}, \tag{10}$$

where the parameter $\eta \geqslant 0$ controls the sparseness of $\mathbf{W}$ and $\|\mathbf{W}\|_{1/2}$ is

$$\|\mathbf{W}\|_{1/2} = \left(\sum_{i=1}^{k}\sum_{j=1}^{n} w_{ij}^{1/2}\right)^2 . \tag{11}$$

Manifold Regularization: Many human motions lie on low-dimensional manifolds that are non-Euclidean [11]. NMF with the $L_{1/2}$ norm sparsity constraint, however, produces a weighting map based on a Euclidean structure in the high-dimensional data space. Thus, the intrinsic and geometric relation between motion features may not be reflected accurately. To remedy this, we incorporate a manifold regularization that respects the intrinsic geometric structure as in [8,12,13]. The manifold regularization favors the local geometric structure and also serves as a smoothness operator, which reduces the interference of noise. Our final objective function incorporating both the manifold regularization and the sparsity constraint can then be given by

$$\mathcal{E}(\mathbf{V},\mathbf{W}) = \frac{1}{2}\|\mathbf{U}-\mathbf{V}\mathbf{W}\|_F^2 + \frac{1}{2}\lambda\mathrm{Tr}(\mathbf{W}\mathbf{L}\mathbf{W}^T) + \eta\|\mathbf{W}\|_{1/2}\,, \tag{12}$$

where λ is a balancing parameter of the manifold regularization, $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix, $\mathbf{Q}$ is a heat kernel weighting, $\mathbf{D}$ is a diagonal matrix where $\mathbf{D}_{jj} = \sum_l \mathbf{Q}_{jl}$, and $\mathbf{L} = \mathbf{D} - \mathbf{Q}$, which is the graph Laplacian.

Minimization: The objective function in Eq. (12) is not convex in both $\mathbf{V}$ and $\mathbf{W}$ and therefore it is not possible to find the global minima. To minimize the objective function, we use a multiplicative iterative method similiar to that used in [13]. Let $\mathbf{\Psi} = [\psi_{mk}]$ and $\mathbf{\Phi} = [\phi_{kn}]$ be the Lagrange multiplier subject to $v_{mk} \geq 0$ and $w_{kn} \geq 0$, respectively. By using the definition of the Frobenius norm, $\|\mathbf{U}\|_F = (\mathrm{Tr}(\mathbf{U}^T\mathbf{U}))^{1/2}$, and matrix calculus, the Lagrangian $\mathcal{L}$ is

$$\begin{aligned}\mathcal{L} = \frac{1}{2}\mathrm{Tr}(\mathbf{U}\mathbf{U}^T) - \mathrm{Tr}(\mathbf{U}\mathbf{W}^T\mathbf{V}^T) + \frac{1}{2}\mathrm{Tr}(\mathbf{V}\mathbf{W}\mathbf{W}^T\mathbf{V}^T)\\ +\frac{\lambda}{2}\mathrm{Tr}(\mathbf{W}\mathbf{L}\mathbf{W}^T) + \mathrm{Tr}(\mathbf{\Psi}\mathbf{V}^T) + \mathrm{Tr}(\mathbf{\Phi}\mathbf{W}^T) + \eta\,\|\mathbf{W}\|_{1/2}\,.\end{aligned} \tag{13}$$

The partial derivatives of $\mathcal{L}$ with respect to $\mathbf{U}$ and $\mathbf{V}$ are

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{V}} &= -\mathbf{U}\mathbf{W}^T + \mathbf{V}\mathbf{W}\mathbf{W}^T + \mathbf{\Psi}\\ \frac{\partial\mathcal{L}}{\partial\mathbf{W}} &= -\mathbf{V}^T\mathbf{U} + \mathbf{V}^T\mathbf{V}\mathbf{W} + \lambda\mathbf{W}\mathbf{L} + \frac{\eta}{2}\mathbf{W}^{-1/2} + \mathbf{\Phi}\end{aligned} \tag{14}$$

By using Karush-Kuhn-Tucker conditions—i.e., $\mathbf{\Psi}_{mk}\mathbf{V}_{mk} = 0$ and $\mathbf{\Phi}_{kn}\mathbf{W}_{kn} = 0$—the final update rule becomes

$$\begin{aligned}&\mathbf{V} \leftarrow \mathbf{V}.*\mathbf{U}\mathbf{W}^T./\mathbf{V}\mathbf{W}\mathbf{W}^T\\ &\mathbf{W} \leftarrow \mathbf{W}.*(\mathbf{V}^T\mathbf{U} + \lambda\mathbf{W}\mathbf{Q})./(\mathbf{V}^T\mathbf{V}\mathbf{W} + \frac{\eta}{2}\mathbf{W}^{-1/2} + \lambda\mathbf{W}\mathbf{D}).\end{aligned} \tag{15}$$

2.4 Spectral Clustering

The non-negative weighting map provides a good measure of regional tissue point similarity. Thus spectral clustering using the weighting map is adopted to determine the cohesive motion patterns as spectral clustering outperforms traditional clustering algorithms such as the K-means algorithm [14].

Once $\mathbf{W}$ is determined from Eq. (15), an affinity matrix $\mathbf{A}$ is constructed:

$$\mathbf{A}(i,j) = \exp\left(-\frac{\|w(i) - w(j)\|_2}{\sigma}\right), \tag{16}$$

where $w(i)$ is the i-th column vector of $\mathbf{W}$ and σ denotes the scale (we set $\sigma = 0.02$ in this work). The column vectors of $\mathbf{W}$ form nodes in the graph, and the similarity $\mathbf{A}$ computed between column vectors of $\mathbf{W}$ form the edge weights. On the affinity matrix, we apply a spectral clustering technique using a normalized cut algorithm [15].

3 Experimental Results

3.1 Experiments Using 2D Data

Since there is no ground truth in our *in vivo* data, we used two 2D datasets to demonstrate the clustering performance of the proposed method. The first

Table 1. Clustering Performance: NMI and AC

NMI (%)	K-means	N-Cut	NMF-K	G-NMF-K	GS-NMF-K	G-NMF-S	Ours
COIL20 (K=20)	73.80%	76.56%	74.36%	87.59%	90.11%	90.24%	**90.63**%
PIE (K=68)	54.40%	77.13%	69.82%	89.93%	89.95%	90.95%	**91.74**%
AC (%)	K-means	N-Cut	NMF-K	G-NMF-K	GS-NMF-K	G-NMF-S	Ours
COIL20 (K=20)	60.48%	66.52%	66.73%	72.22%	83.75%	84.58%	**85.00**%
PIE (K=68)	23.91%	65.91%	66.21%	79.3%	79.93%	80.60%	**84.31**%

dataset is the COIL20 image library, which contains 20 classes (32×32 gray scale images of 20 objects). The second dataset is the CMU PIE face database, which has 68 classes (32×32 gray scale face images of 68 persons). In order to compare the performance of the different algorithms, we used a K-means clustering method (K-means), a normalized cut method (N-Cut) [15], standard NMF with K-means clustering (NMF-K), graph-regularized NMF with K-means clustering (G-NMF-K) [8], graph-regularized NMF with spectral clustering (G-NMF-S), graph-regularized sparse NMF with K-means clustering (GS-NMF-K), and our method (GS-NMF-S). Two metrics, the Normalized Mutual Information (NMI) and the accuracy (AC), were used to measure the clustering performance as used in [8]. Table 1 lists the NMI and AC values, demonstrating that the proposed method outperformed other methods. We also compared the L1/2 and L1 norms experimentally, and the L1/2 norm had slightly better results.

3.2 Experiments Using *In Vivo* Tongue Data

We also tested our method using a simple non-speech protrusion task and a speech task: “a souk”. Four subjects said “a souk” and one subject performed the protrusion task. All MRI scanning was performed on a Siemens 3.0 T Tim Treo system (Siemens Medical Solutions, Malvern, PA) with 12-channel head and 4-channel neck coil. The tMRI datasets were collected using Magnitude Image CSPAMM Reconstructed images [16]. The datasets had a 1 second duration, 26 time-frames per sec, 6 mm thick slices, 6 mm tag separation, 1.875 mm in-plane resolution. The field-of-view was 24 cm. We acquired 2D orthogonal stacks of tMRI and used harmonic phase (HARP) to track internal tissue points. The incompressible deformation estimation algorithm was then used to combine 2D tracking data to produce the 3D tracking result [17].

Fig. 1 shows the protrusion task. The outer tongue layer expands forward and upward (but not backward), and the region near the jaw has little motion. Functional units, based on magnitude and angle, have been extracted for two (Fig. 1(b)) and three clusters (Fig. 1(c)). Fig. 1(b) is a good representation of forward protrusion (blue) vs no motion (red), but the three cluster output introduces noise, suggesting there are only two clusters, or functional units.

Fig. 2 shows the motion from /*s*/ to /*u*/ during the word “a souk”. The functional units were determined using our method for two clusters (Fig. 2(b)) and three clusters (Fig. 2(c)), respectively. Note that the three clusters better

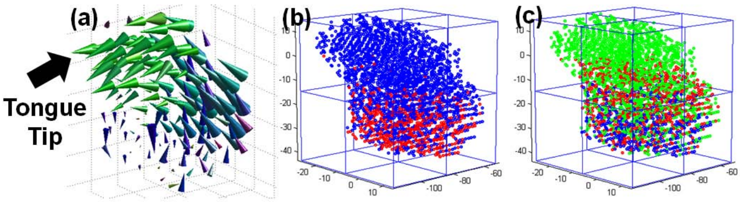

Fig. 1. Illustration of protrusion showing (a) 3D displacement field, (b) functional units (2 clusters), (c) functional units (3 clusters)

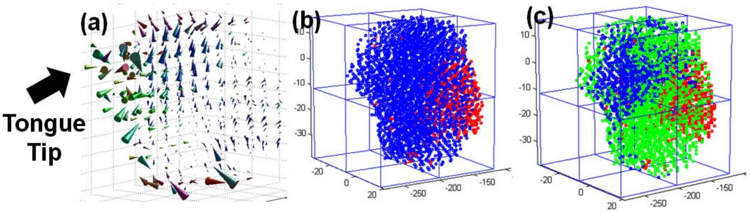

Fig. 2. Illustration of motion from /*s*/ to /*u*/ showing (a) 3D displacement field, (b) functional units (2 clusters), (c) functional units (3 clusters)

represent the motions of the tongue. These motions include backward motion of the tongue tip (blue), upward motion of the tongue body (green), and forward motion of the posterior tongue (red).

4 Discussion and Conclusion

In this work, inspired by recent advances in sparse NMF and manifold learning, we presented a novel method for determining functional units from tMRI. Unlike previous algorithms, this proposed work aims at identifying the internal, coherent manifold structure of high-dimensional motion data to determine functional units. The contributions of this work are two-fold. In an NMF framework, we formulate a new clustering problem, that of a learning latent weighting map as well as spectral clustering and we give an efficient algorithm to solve this problem. Our method performed better than K-means, N-Cut, NMF-K, G-NMF-K, GS-NMF-K, and G-NMF-S using 2D data. In a tongue motion analysis context, we define functional units from tMRI, which opens new vistas to study speech production. The identified functional units are visually assessed and further studies using biomechanical stimulations are needed to co-validate our findings due to the lack of ground truth in *in vivo* data. The proposed method gives a principled method for defining subject/task-specific functional units, which can be potentially used to elucidate speech-related disorders.

Acknowledgements. We thank the reviewers for their comments. This work is supported by NIH/NCI 5R01CA133015 and NIH/NIDCD K99 DC012575.

References

1. Green, J.R.: Tongue-surface movement patterns during speech and swallowing. The Journal of the Acoustical Society of America 113(5), 2820–2833 (2003)
2. Stone, M., Davis, E.P., Douglas, A.S., Aiver, M.N., Gullapalli, R., Levine, W.S., Lundberg, A.J.: Modeling tongue surface contours from cine-MRI images. Journal of Speech, Language, and Hearing Research 44(5), 1026 (2001)
3. Bresch, E., Kim, Y.C., Nayak, K., Byrd, D., Narayanan, S.: Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging. IEEE Signal Processing Magazine 25(3), 123–132 (2008)
4. Parthasarathy, V., Prince, J.L., Stone, M., Murano, E.Z., NessAiver, M.: Measuring tongue motion from tagged cine-MRI using harmonic phase (HARP) processing. The Journal of the Acoustical Society of America 121(1), 491–504 (2007)
5. Stone, M., Epstein, M.A., Iskarous, K.: Functional segments in tongue movement. Clinical Linguistics & Phonetics 18(6-8), 507–521 (2004)
6. Ramanarayanan, V., Goldstein, L., Narayanan, S.S.: Spatio-temporal articulatory movement primitives during speech production: Extraction, interpretation, and validation. JASA 134(2), 1378–1394 (2013)
7. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)
8. Cai, D., He, X., Han, J., Huang, T.S.: Graph regularized nonnegative matrix factorization for data representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(8), 1548–1560 (2011)
9. Cheriyadat, A.M., Radke, R.J.: Non-negative matrix factorization of partial track data for motion segmentation. In: IEEE 12th International Conference on Computer Vision, pp. 865–872 (2009)
10. Qian, Y., Jia, S., Zhou, J., Robles-Kelly, A.: Hyperspectral unmixing via sparsity-constrained nonnegative matrix factorization. IEEE Transactions on Geoscience and Remote Sensing 49(11), 4282–4297 (2011)
11. Elgammal, A., Lee, C.S.: The role of manifold learning in human motion analysis. In: Rosenhahn, B., Klette, R., Metaxas, D. (eds.) Human Motion. Computational Imaging and Vision, vol. 36, pp. 25–56. Springer, Netherlands (2008)
12. Yang, S., Hou, C., Zhang, C., Wu, Y.: Robust non-negative matrix factorization via joint sparse and graph regularization for transfer learning. Neural Computing and Applications 23(2), 541–559 (2013)
13. Lu, X., Wu, H., Yuan, Y., Yan, P., Li, X.: Manifold regularized sparse NMF for hyperspectral unmixing. IEEE Transactions on Geoscience and Remote Sensing 51(5), 2815–2826 (2013)
14. Von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing 17(4), 395–416 (2007)
15. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888–905 (2000)
16. NessAiver, M., Prince, J.L.: Magnitude image CSPAMM reconstruction (MICSR). Magnetic Resonance in Medicine 50(2), 331–342 (2003)
17. Xing, F., Woo, J., Murano, E.Z., Lee, J., Stone, M., Prince, J.L.: 3D tongue motion from tagged and cine MR images. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part III. LNCS, vol. 8151, pp. 41–48. Springer, Heidelberg (2013)