Selecting Scales by Multiple Kernel Learning for Shape Diffusion Analysis

Umberto Castellani[1⋆], Aydın Ulaş[1], and Vittorio Murino[1,2],
Marcella Bellani[3], Gianluca Rambaldelli[3], Michele Tansella[3], Paolo
Brambilla[3,4]

[1] Department of Computer Science, University of Verona, Italy
[2] Istituto Italiano di Tecnologia, Genova, Italy
[3] Department of Medicine and Public Health, University of Verona, Italy
[4] ICBN Center, University of Udine and Verona, Italy

Abstract. Brain morphological abnormalities can typically be detected
by advanced geometrical shape analysis techniques. Recently, shape dif-
fusion methods have proved to be very effective in providing useful de-
scriptions for brain classification purposes. In particular, they allow the
analysis of such shapes at multiple scales, but the selection of the correct
range of scales remains an open issue heavily affecting the performance
of methods, and it needs to be estimated adaptively for different classes
of shapes. In this paper, we focus on the diffusion scale selection in order
to define a robust shape descriptor for brain classification. To this end,
geometric features are extracted for each scale and the best feature com-
bination is selected by employing *multiple kernel learning* (MKL). In the
presented experiments, we compare the shape of Thalamic regions in or-
der to discriminate between normal subjects and schizophrenic patients.
We demonstrate that MKL allows to obtain classifiers which are more
accurate with respect to other competing algorithms for schizophrenia
detection. Moreover, using the weights computed by the MKL algorithm,
we can select at which scale the features are more effective for schizophre-
nia classification.

Keywords: multiple kernel learning, schizophrenia, heat kernel, spec-
tral shape analysis, support vector machines

1 Introduction

Recent advances in geometric shape analysis have led to a larger diffusion of
computational anatomy methods, aimed at characterizing or modeling the mor-
phological variations of biological shapes. One of the typical applications is ana-
lyzing the anatomy of organs, known as being possibly affected by abnormalities
due to a certain disease, of several persons in order to discriminate between
normal and pathological subjects [11,1]. To this aim, effective shape analysis

---

⋆ Corresponding author.

techniques are crucial to extract geometric features with high discriminant properties. A wide class of methods are based on the encoding of the deformation which aligns a pair of subjects, but such an approach requires the solution of a complex problem due to the non-linear registration between different shapes. More recently, new methods have been proposed to encode the shape geometric properties into a *descriptor* which compactly represents the whole shape, and performing the comparison by computing similarities in the descriptor space without any registration procedure. Among the shape analysis methods, diffusion geometry approaches are very promising since they are able to capture *intrinsic* characteristic of the shape. More specifically local geometric properties are encoded by the so-called *Heat Kernel* [16] which exploits heat diffusion characteristics at a given scale. The general idea consists of gaining information about the neighborhood of a point on the shape by recording the dissipation of heat over time from that point onto the rest of the shape. The fixed time is very important since it allows to capture different kinds of information: *local* shape characteristics are highlighted through the behavior of heat diffusion over short time periods, and, conversely, *global* shape properties are observed while considering longer periods [16,10]. So doing, simply varying a single parameter (the time), it is possible to characterize the properties of a shape at different scales. In particular, the so called *Heat Kernel Signature*(HKS) [16] has been proposed to encode simultaneously the contribution of local features for a fixed set of scales into a single shape descriptor. This general approach has been succesfully applied for object retrieval [4] and brain classification [6]. However, the choice of the range of the time periods to be evaluated (i.e., the *scales*) is critical and depends on the considered shape. In fact, for a particular shape, some scales may be highly discriminative, while some other scales should encode useless information. In this paper, we propose a new approach for integrating and selecting the contribution of geometric features collected at different scales by utilzing a Multiple Kernel Learning (MKL) approach. In general, MKL algorithms can learn a weighted combination of different kernel functions able to exploit information coming from multiple sources. In our case, the different sources are the features extracted at different scales. Therefore, several kernels are computed (i.e., one kernel per scale) and a set of weights are estimated for the kernel combination. In this fashion, we can choose the most discriminative scales by selecting those associated to the highest weights, and viceversa. Moreover, kernel combination leads to a new similarity measure which increases the classification accuracy. It is important to note that in our appraoch we aim at selecting the best shape characteristics for classification purposes, hence, our selection is driven by the performance of a Support Vector Machine (SVM) classifier. We have applied our method for brain classification in schizophrenic subjects: we have adopted a Region of Interest (ROI)-based method by analysing the shape of the Thalamic region, employing a *volumetric*-heat kernel computed for each voxel of the MRI scan at different scales, as described in our previous work [6]. This paper improves [6] for both methodological aspects, by proposing the automatic scale selection procedure and promising results. The rest of the paper is organized as

follows. In Section 2, the basics on shape diffusion procedures are reported. Section 3 describes the Multiple Kernel Lerning strategy, and the proposed method is reported in Section 4. Results are shown in Section 5 and conclusions are finally drawn in Section 6.

2   Shape analysis by heat diffusion

Considering a shape M as a compact Riemannian manifold [5], the heat diffusion on shape[5] is defined by the *heat* equation:

$$(\Delta_M + \frac{\partial}{\partial t})u(t, \boldsymbol{m}) = 0;\tag{1}$$

where u is the distribution of heat on the surface, $\boldsymbol{m} \in M$, Δ_M is the *Laplace-Beltrami* operator which, for compact spaces, has discrete eigendecomposition of the form $\Delta_M = \lambda_i \phi_i$. In this way, the *heat kernel* has the following eigendecomposition:

$$h_t(\boldsymbol{m}, \boldsymbol{m}') = \sum_{i=0}^{\infty} e^{-\lambda_i t} \phi_i(\boldsymbol{m}) \phi_i(\boldsymbol{m}'),\tag{2}$$

where λ_i and ϕ_i are the ith eigenvalue and the ith eigenfunction of the Laplace-Beltrami operator, respectively. The heat kernel $h_t(\boldsymbol{m}, \boldsymbol{m}')$ is the solution of the heat equation with initial point heat source in $\boldsymbol{m}$ at time $t = 0$, and heat value in ending point $\boldsymbol{m}' \in M$ after time t. The heat kernel is *isometric invariant*, it is *informative*, and *stable* [16].

In the case of volumetric representations, the volume is sampled by a regular Cartesian grid composed by voxels, which allows the use of standard Laplacian in R^3 as the Laplace-Beltrami operator. We use finite differences to evaluate the second derivative in each direction of the volume. The heat kernel on volumes is invariant to volume isometries, in which shortest paths between points inside the shape do not change. Note that in real applications exact volume isometries are limited to the set of rigid transformations [15], however, also non-rigid deformations can faithfully be modelled as approximated volume isometries in practice. It is also worth noting that, as observed in [16,15], for small t the autodiffusion heat kernel $h_t(\boldsymbol{m}, \boldsymbol{m})$ of a point $\boldsymbol{m}$ with itself is directly related to the *scalar curvature* $s(\boldsymbol{m})$ [15]. More formally:

$$h_t(\boldsymbol{m}, \boldsymbol{m}) = (4\pi t)^{-3/2}(1 + \frac{1}{6}s(\boldsymbol{m})).\tag{3}$$

In practice, Equation 3 states that the heat tends to diffuse slower at points with positive curvature, and viceversa. This gives an intuitive explanation about the geometric properties of $h_t(\boldsymbol{m}, \boldsymbol{m})$, and suggests the idea of using it to build a shape descriptor [16].

---

[5] In this section, we borrow the notation from [16,5]

3   Multiple Kernel Learning

The main idea behind kernel methods [17] is to transform the input feature space to another space (eventually with a larger dimension) where the classes are linearly separable. In particular, by employing the SVM classifier, the discriminant function after the training phase becomes $f(x) = \langle \boldsymbol{w}, \Phi(\boldsymbol{x}) \rangle + b$, where $\boldsymbol{w}$ and b are the parameters of the hyperplane which separates two classes, and $\Phi(\cdot)$ is the mapping function. Using the dual formulation and the kernel trick, one does not have to define this mapping function explicitly and the discriminant function can be written as

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b \qquad (4)$$

where $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle$ is the kernel function that calculates a similarity metric between data instances.

More recently, MKL methods have been proposed [3,13] for learning a combination k_η of several kernels:

$$k\eta(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\eta}) = f_\eta(\{k_m(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m)_{m=1}^{P}\}; \boldsymbol{\eta}) \qquad (5)$$

where the combination function f_η forms a single kernel from P base kernels using the parameters vector (i.e., weights) $\boldsymbol{\eta}$. Such new kernel must be a valid kernel[6] [9] and can be plugged in Equation 4 for classification purposes. Different kernel functions correspond to different notions of similarity and instead of searching which works best, the MKL method does the picking for us, or may use a combination of kernels. MKL also allows us to combine different representations, possibly coming from different sources or modalities.

There is significant work on the theory and application of MKL, and most of the proposed algorithms differ among them by the optimization method employed to estimate the weights and by the used combination rule [3,13,14]. In this paper we focus on *linear*-MKL methods [3,13], whose general formulation is defined as:

$$k\eta(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\eta}) = \sum_{m=1}^{P} \eta_m k_m(\boldsymbol{x}_i^m, \boldsymbol{x}_j^m) \qquad (6)$$

with $\eta_m \in \mathbb{R}$. As a simplest combination approach, the so called *fixed rules* [9] use the combination function in Eq. (6) with all weights equally set to $\boldsymbol{\eta}_m = 1$. Similarly, the *mean*-rule takes the mean of the kernels by setting all $\boldsymbol{\eta}_m = 1/P$. Indeed, in the most general case the weights $\boldsymbol{\eta}_m$ are automatically estimated by a *learning by example* approach. More specifically, MKL methods search for a combination of kernels that maximizes a generalized performance measure (i.e, *maximum margin* classification errors [9]). To this aim, in the training phase, both MKL weights and SVM parameters are simultaneously estimated within the same optimization problem.

---

[6] The validity of the kernel depends by the combination function.

4   The Proposed Method

The proposed method can be summarized in the following main steps:

1. MRI data collection.
2. Feature extraction at multiple scales.
3. Learning weights and classifier by MKL.
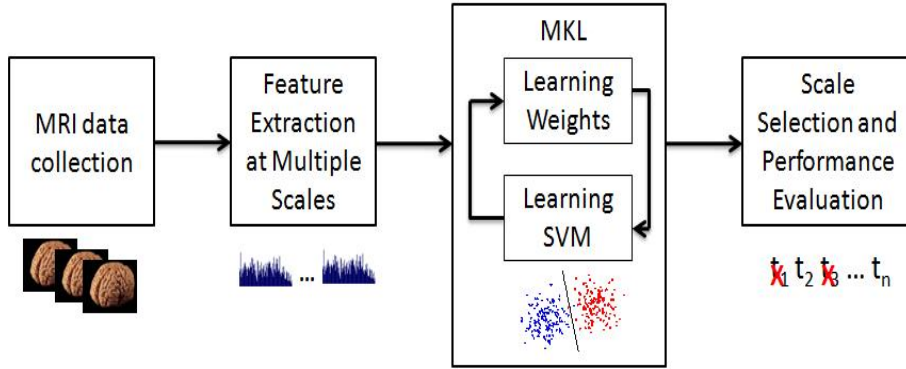4. Scale selection and performance evaluation.

Fig. 1. General scheme of the proposed method.

MRI data collection. In order to employ a *learning-by-example* approach, we need a collection of samples for both healthy subjects and patients. Source data are MRI scans where shape information can be provided in terms of volumetric data.

Feature extraction at multiple scales. According to the shape diffusion analysis described in Section 2, for each subject geometric features are extracted at multiple scales: a set of time values $(t_1, t_2, \cdots, t_n)$ are defined, and the autodiffusion value is computed for each voxel $\boldsymbol{m}$, leading to:

$$H_{t_i}(M) = \{h_{t_i}(\boldsymbol{m}, \boldsymbol{m}), \forall \boldsymbol{m} \in M\}.$$

Then, such values are accumulated into a histogram $r_i = hist(H_{t_i}(M))$. In this manner, we obtain a set of n sources of shape representation $\{r_1, \cdots, r_n\}$, each one encoding the global shape at a certain scale. The number of bins for each histogram is chosen as 100.

Learning weights and classifier by MKL. The contribution of geometric features extracted at each scale are combined by employing the MKL strategy as described in 3. Each shape representation r_i is associated to a kernel k_m by leading to $n = P$ kernels. Indeed, both the weights $(\eta_1 \cdots \eta_P)$ and the SVM parameters are estimated. In order to obtain the best classification accuracy according to the *max-margin* paradigm an *alternating* approach is used between the optimization of kernel weights and the optimization of the SVM classifier. In each step, given the current solution of kernel weights, MKL solves a standard SVM optimization problem with the combined kernel. Then, a specific procedure is applied to update the kernel weights.

Scale selection and performance evaluation. Once the MKL procedure is completed, we obtain a two-fold advantage: i) we can select the best scale contributions by keeping only the scales associated to the highest weighs, and ii) we can compose a new kernel from the weighted contributions of the best scales, which can be evaluated for classification purposes.

5   Experiments

This section is organized in the following parts: i) data gathering, ii) experimental methodology, iii) results, and iv) discussion.

5.1   Data Gathering

Quantitative data collection and processing in MRI-based research implies to face several methodological issues to minimize biases and distortions. The standard approach to deal with these issues is following well-established guidelines dictated by international organizations, such as the World Health Organization (WHO), or codified by respected institutions, such as leading universities. All patients received a diagnosis of schizophrenia according to the criteria of the Diagnostic and Statistical Manual of Mental Disorders [2]. In this work, we employ a ROI-based approach [11], so only a well defined brain subpart has been considered in this study. More specifically, we focus our analysis on the left-Thalamus whose abnormal activity has been already investigated in schizophrenia[8]. ROIs have been manually traced by experts, according to well defined medical protocols. The data set used in this work is composed by MRI brain scans of 30 patients affected by schizophrenia and 30 healthy control subjects.

5.2   Experimental protocol

In our experiments, we apply leave-one-out (LOO) cross-validation to assess the performance of the technique. Since LOO is used as the cross validation technique, we do not report standard deviations or variances. We compare our results using k-fold paired t-test at $p = 0.05$. We collect geometric features at 11 scales generating different shape representations $r_{01}, \cdots, r_{11}$. In practice, each

representation r_i is a feature vector x_i which is plugged in the MKL framework. We employ the dot product as basic kernel function (i.e., linear kernel) since it avoids the estimation of free kernel parameters. Different strategies to combine the different shape representations have also been evaluated:

- Single Best Kernel (`Single-best`): an SVM is trained separately per each representation. Therefore, the performance of the classification are evaluated separately at each scale. So doing, we can evaluate the independent contributions coming from the different sources of information and select the best one.
- Feature concatenation (`SVM-con`): the contributions coming from the different sources are concatenated into a single feature vector. Then, a single SVM is employed for classification [7].
- Rule-based MKL (`RBMKL`): as baseline MKL approach, the so called rule-based method is evaluated: the kernels computed at each scale are combined by simply taking their average (i.e., $\forall m, \boldsymbol{\eta_m} = 1/P$).
- Simple MKL (`SimpleMKL`): a simple but effective MKL algorithm is employed [14] by addressing the MKL problem through a weighted 2-norm regularization formulation with additional constraint on the weights that encourages sparse kernel combination. It is a popular approach and its code is publicly available[8].
- Group Lasso MKL (`GLMKL`): it denotes the group Lasso-based MKL algorithms proposed by [12,18]. A closed form solution for optimazing the kernel weights based on the equivalence between group-lasso and MKL is proposed. In our implementation, we used l_1-norm on the kernel weights and learned a convex combination of the kernels.

5.3   Results

The first evaluation scores are shown in Table 1, which reports the single-best kernel accuracies for all feature representations. We can observe that the best performance is obtained at 78.33 % using `r02` which is shown as bold face in the table. The entries marked with "*" show the accuracies which are statistically significantly less accurate than the best algorithm using k-fold paired t-test at $p = 0.05$.

Table 1. Single-kernel SVM accuracies.

r01	r02	r03	r04	r05	r06	r07	r08	r09	r10	r11
75.00	**78.33**	76.67	76.67	73.33	*66.67	68.33	70.00	76.67	71.67	70.00

Second, concatenating the features in a single vector leads to 83.33 % accuracy.

---

[7] We use LIBSVM software [7] to train the SVM.

[8] http://asi.insa-rouen.fr

Third, using the proposed three different MKL algorithms, we combined the eleven kernels by introducing the weights $\boldsymbol{\eta_m}$. Table 2 reports the results of the best single-kernel SVM, the accuracy of the concatenated feature set, and the three MKL-based algorithms trained. The values in parantheses show the percentage of controls classified as schizophrenia and the percentage of patients classified as healthy respectively. We achieve an accuracy of 86.67%, reached by combining eleven kernels with the `SimpleMKL` approach. This result is better than all other MKL settings and single-kernel SVMs. Further, `GLMKL` achieves 85% accuracy which is still higher than that reached by the feature concatenation method. We can also note that we cannot overcome `SVM-con` when we use `RBMKL`, as the latter gives equal weight to each kernel. In fact, if there are inaccurate representations in the given set, the overall mean combination accuracy may be less of that reached using the single best. Conversely, when the weights are automatically estimated, such as in `SimpleMKL` and `GLMKL` the selection of the most reliable information is carried out by the MKL procedure and the overal performance improves.

Table 2. MKL accuracies (false positives and negatives are reported in brackets).

Single-best	SVM-con	RBMKL	SimpleMKL	GLMKL
*78.33 (10, 11.6)	83.33 (8.3, 8.3)	*81.67 (10, 8.3)	**86.67** (6.6, 6.6)	85.00 (8.3, 6.6)

In Figure 2, we plotted the weights of MKL for both `SimpleMKL` and `GLMKL` algorithms. Note that the estimated weights are coherent in the two algorithms. As expected, the best representation is `r02`, which has the highest weights. Although the other representations with high weights (`r08`, `r11` and `r05`) do not provide much accurate single-kernel SVMs results, their contributions to the overall accuracy in the combination is higher than those given by the other kernels. This demonstrates that when considering combinations, even a representation which does not lead to very precise results may contribute to raise the overall combination accuracy. Moreover, we can also deduce that these four representations are the most useful in discriminating between healthy and schizophrenic subjects, and we may focus the attention on these properties only.

Using this information, we also performed the above pipeline using only these four representations, and we can observe the results in Table 3. Using this subset, we get the highest accuracy with `SimpleMKL`[9], reaching 88.33% of accuracy. We can also observe an increase in `RBMKL`.

---

[9] Note that in principle the same result should have been obtained automatically from MKL algorithms on all representations. In practice, this is not the case in our experiment due to the fact that the estimated solution is trapped into a local minimum.
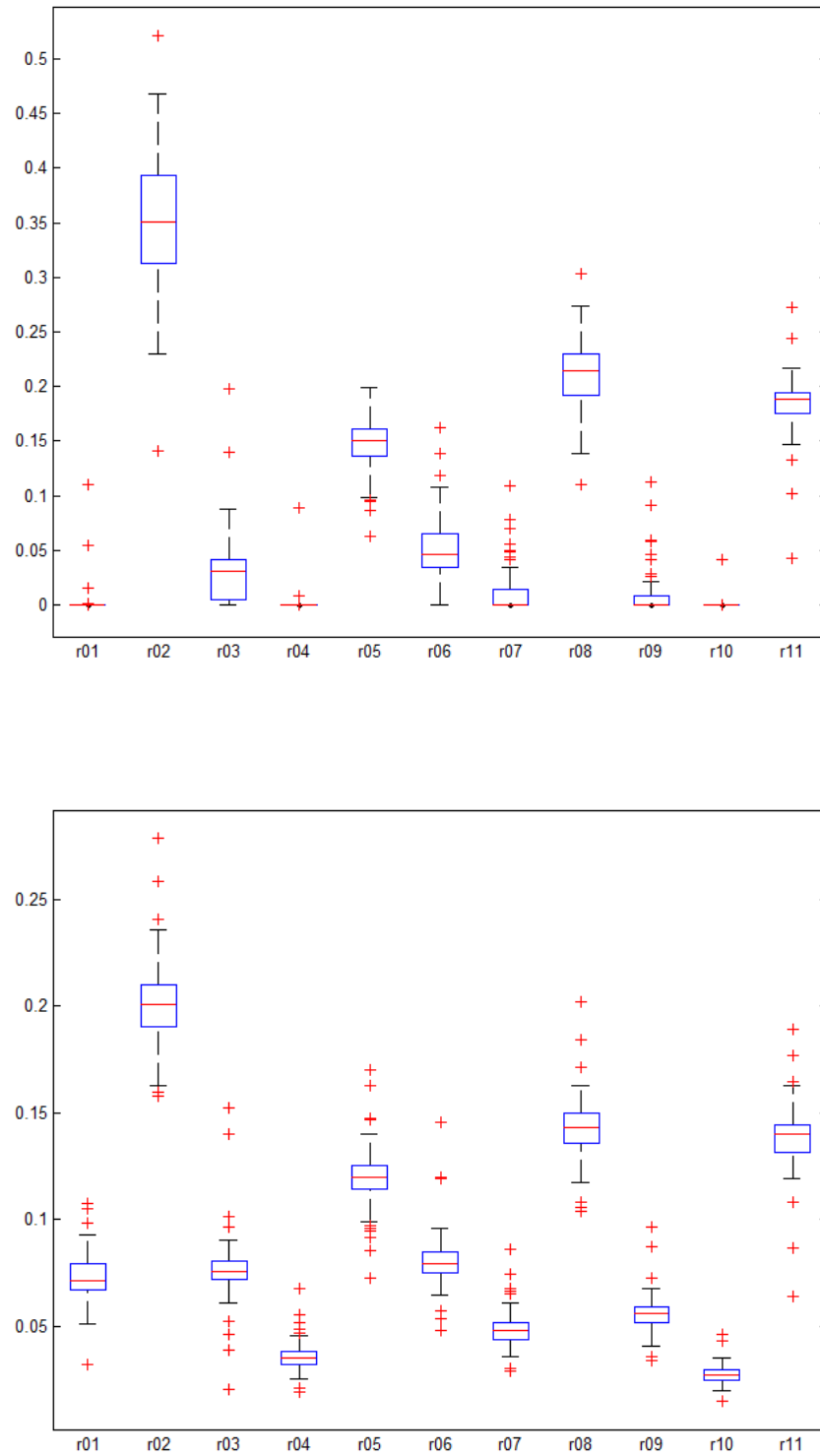
Fig. 2. Combination weights in MKL using the linear kernel. Top: using `SimpleMKL`, Bottom: using `GLMKL`.

Table 3. MKL accuracies on the selected subset of representations (false positives and negatives are reported in brackets).

SVM	SVM-con	RBMKL	SimpleMKL	GLMKL
*78.33 (10, 11.6)	*83.33 (6.6, 10)	*83.33 (6.6, 10)	**88.33** (6.6, 5)	85.00 (6.6, 8.3)

5.4 Discussion

In this work, we have shown in general that MKL algorithms perform better than both single-best kernel SVMs and feature concatenation strategies. We have also observed that RBMKL (which does not compute weights while combining kernels) does not outperform the feature concatenatenation approach. Conversely, when the kernel combination is carried out by estimating proper weights, a drastic improvement is instead obtained. The kernel weights also allow us to extract useful information: it is interesting to observe that, for both MKL algorithms with the highest accuracy, four representations have the maximum effect (i.e., the highest weights), i.e., r02, r08, r11, and r05, with r02 being the best single-kernel. We use this information to select a smaller number of representations to reduce the costs of the feature extraction phase. Finally, we can also observe that by using such subset we can reach the best accuracy overall.

6   Conclusions

In this paper, we focus on scale selection for anatomical shape characterization. By employing a shape diffusion approach, we extract several shape descriptors at different scales in order to discriminate between healthy subjects and patients affected by schizophrenia. We have shown that machine learning techniques can be useful to improve the shape analysis in these (biomedical) contexts. We propose a Multiple Kernel Learning algorithm for the automatic estimation of the best feature representation for classification purposes. In this way, being driven by the training data, we are able to choose the scales of the heat kernel which are more suitable to describe our kind of shapes. In particular, in our experiments addressing the Thalamic region classification, we have shown that both small and high scales are crucial. Actually, the best accuracy is observed at r02 for which very local information are collected from the shape. Nevertheless, when also higher scales are considered the performance is further improved, meaning that also global shape information is relevant.

Acknowledgements

References

1. Agarwal, N., Port, J.D., Bazzocchi, M., Renshaw, P.F.: Update on the use of MR for assessment and diagnosis of psychiatric diseases. Radiology 255(1), 23–41 (2010)
2. American Psychiatric Association: Diagnostic and statistical manual of mental disorders, DSM-IV. Washington DC, 4th edn. (1994)
3. Bach, F.R., Lanckriet, G.R.G., Jordan, M.I.: Multiple kernel learning, conic duality, and the smo algorithm. In: Proceedings of the twenty-first international conference on Machine learning, ICML '04. pp. 41–48 (2004)
4. Bronstein, A.M., Bronstein, M.M., Ovsjanikov, M., Guibas, L.J.: Shape Google: geometric words and expressions for invariant shape retrieval. ACM Transaction on Graphics 30(1), 1–20 (2011)
5. Bronstein, A.M., Bronstein, M.M., Ovsjanikov, M., Guibas, L.J.: Shape recognition with spectral distances. IEEE Trans. Pattern Analysis and Machine Intelligence 33(5), 1065–1071 (2011)
6. Castellani, U., Mirtuono, P., Murino, V., Bellani, M., Rambaldelli, G., Tansella, M., Brambilla, P.: A new shape diffusion descriptor for brain classification. In: Medical Image Comp. Computer-Assisted Intervention (MICCAI) (2011)
7. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), `http://www.csie.ntu.edu.tw/$\sim$cjlin/libsvm`
8. Corradi-DellAcqua, C., Tomelleri, L., Bellani, M., Rambaldelli, G., Cerini, R., Pozzi-Mucelli, R., Balestrieri, M., Tansella, M., Brambilla, P.: Thalamic-insular dysconnectivity in schizophrenia: Evidence from structural equation modeling. Human Brain Mapping p. in press (2011)
9. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and other Kernel-based Learning Methods. Cambridge University Press (2000)
10. Gebal, K., Baerentzen, J.A., Aanaes, H., Larsen, R.: Shape analysis using the auto diffusion function. In: In SGP (2009)
11. Giuliani, N.R., Calhouna, V.D., Pearlson, G.D., Francis, A., Buchanan, R.W.: Voxel-based morphometry versus region of interest: a comparison of two methods for analyzing gray matter differences in schizophrenia. Schizophrenia Research 74(2–3), 135–147 (2005)
12. Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A.: l_p-norm multiple kernel learning. Journal of Machine Learning Research 12, 953–997 (2011)
13. Lanckriet, G.R.G., Cristianini, N., Bartlett, P., Ghaoui, L.E., Jordan, M.I.: Learning the kernel matrix with semidefinite programming. Journal of Machine Learning Research 5, 27–72 (December 2004)
14. Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: SimpleMKL. Journal of Machine Learning Research 9, 2491–2521 (2008)
15. Raviv, D., Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Volumetric heat kernel signatures. In: Workshop on 3D Object Retrieval (2010)
16. Sun, J., Ovsjanikov, M., Guibas, L.: A concise and provably informative multi-scale signature based on heat diffusion. In: Proceedings of the Symposium on Geometry Processing. pp. 1383–1392 (2009)
17. Vapnik, V.N.: Statistical learning theory. John Wiley and Sons (1998)
18. Xu, Z., Jin, R., Yang, H., King, I., Lyu, M.R.: Simple and efficient multiple kernel learning by group Lasso. In: Proceedings of the 27th International Conference on Machine Learning, ICML '10. pp. 1175–1182 (2010)