SEMI-SUPERVISED HANDWRITTEN DIGIT RECOGNITION USING VERY FEW LABELED DATA

Steven Van Vaerenbergh, Ignacio Santamaria *

Department of Communications Engineering
University of Cantabria
Santander, Spain

Paolo Emilio Barbano

Dept. of Applied Math. and Theoretical Physics
University of Cambridge
Cambridge, UK

ABSTRACT

We propose a novel semi-supervised classifier for handwritten digit recognition problems that is based on the assumption that any digit can be obtained as a slight transformation of another sufficiently close digit. Given a number of labeled and unlabeled images, it is possible to determine the class membership of each unlabeled image by creating a sequence of such image transformations that connect it, through other unlabeled images, to a labeled image. In order to measure the total transformation, a robust and reliable metric of the path length is proposed, which combines a local dissimilarity between consecutive images along the path with a global connectivity-based metric. For the local dissimilarity we use a symmetrized version of the zero-order image deformation model (IDM) proposed by Keysers et al. in [1]. For the global distance we use a connectivity-based metric proposed by Chapelle and Zien in [2]. Experimental results on the MNIST benchmark indicate that the proposed classifier outperforms current state-of-the-art techniques, especially when very few labeled patterns are available.

Index Terms— Semi-supervised classification, handwritten character recognition, connectivity, deformation models.

1. INTRODUCTION

For many classification problems, obtaining labeled training data is a time-consuming and expensive task, whereas large unlabeled data sets are typically available. This is particularly true in many problems involving high-dimensional data, e.g., handwritten digit recognition [1] or protein classification [3]. Semi-supervised classification algorithms exploit this setting by making efficient use of both labeled and unlabeled data [4, 5]. The success of these techniques relies mainly on two key assumptions: i) two data points are similar to each other if they are connected by a high-density region (cluster assumption), and ii) the input data lie on or close to a low-dimensional manifold (manifold assumption).

Most semi-supervised classifiers start by constructing an undirected weighted graph on the labeled and unlabeled data points, where the edge weights measure the pairwise similarities, and then apply different approaches to design a global classifying function with some desirable properties (e.g., smoothness, robustness, etc.). Typically, they use a local dissimilarity measure based on the Euclidean distance (see for instance [4, 6, 7, 8, 9]), regardless of the particular application considered. Their emphasis is on how a suitable global metric or function should be estimated from that graph, and to this end they proposed quite sophisticated methods.

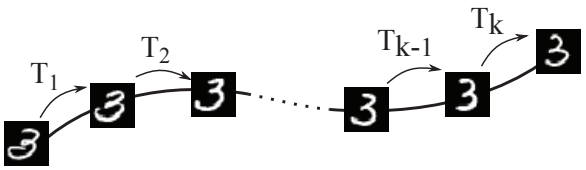

Fig. 1. By applying a short sequence of small transformations it is possible to transform any pattern into another pattern of the same class. All patterns of the sequence belong to the same low-dimensional manifold.

Nevertheless, it is well-known that all pairwise Euclidean distances seem to be similar in high-dimensional data sets. This observation is sometimes referred to as the "concentration phenomenon" in the pattern recognition literature [10]. The limitations of the Euclidean distance can be illustrated with a simple example. Suppose we are given two images that are identical except for the fact that one image is shifted a few pixels to the right. Although these images are almost identical, the Euclidean distance between vector-representations of these images will indicate a high dissimilarity. Hence, in image classification problems it seems necessary to use distances that are invariant to certain transformations of the input.

In this paper we translate most of the complexity to the computation of the local dissimilarity measure. This metric should be problem-dependent to better characterize the data manifold structure at a local scale. In particular, it should be less sensitive to small image transformations (e.g., translations, rotations, scaling and other small deformations) and to noise. We use a symmetrized version of the zero-order image distortion model (IDM) as a local dissimilarity measure, originally proposed by Keysers et al. in [1] for supervised classification, because of its good tradeoff between computational simplicity and matching flexibility. This allows us to simplify the global metric as a shortest path based metric, which can be implemented using Dijkstra's algorithm. As we will show, this conceptually simple procedure provides very good results in problems that involve series of spatially related images.

The remainder of the paper is organized as follows. In Section 2 we review the semi-supervised classification setting and describe the main assumptions on which the proposed method is based. Section 3 describes the local and global similarity measures, which form the basis of the proposed semi-supervised classifier and Section 4 illustrates the obtained results on the MNIST database. Finally, we summarize the main conclusions of this work in Section 5.

*This work was supported by MICINN (Spanish Ministry for Science and Innovation) under grants TEC2007-68020-C04-02 TCM (MULTIMIMO), TEC2010-19545-C04-03 (COSIMA) and CONSOLIDER-INGENIO 2010 CSD2008-00010 (COMONSENS).

2. PROBLEM SETTING

Consider a classification problem with N classes $\{\mathcal{C}_1, \ldots, \mathcal{C}_N\}$. In a semi-supervised classification setting, we are given a labeled data set containing m patterns, $\mathcal{X}_l = \{\mathbf{x}_i \in \mathbb{R}^d \mid i = 1, \ldots, m\}$ with class labels $\mathcal{Y}_l = \{y_i \in \{1, 2, \ldots, N\} \mid i = 1, \ldots, m\}$, and an unlabeled data set containing n patterns, $\mathcal{X}_u = \{\mathbf{x}_j \in \mathbb{R}^d \mid j = m+1, \ldots, m+n\}$. We assume that all input patterns $\mathcal{X}_l \bigcup \mathcal{X}_u$ have been drawn independently and identically distributed (i.i.d.) from some unknown marginal data distribution $P(\mathbf{x})$. This is the conventional setting assumed in most semi-supervised classification techniques described in the literature [4]. Furthermore, we consider semi-supervised classification problems where m is very small compared to the total number of available unlabeled patterns, n.

The proposed approach relies on the following assumptions:

Assumption 1 *If two patterns $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class ($y_i = y_j$), there exists a sequence of k consecutive geometric transformations*

$$\mathbf{x}_i = T_k \circ T_{k-1} \circ \cdots \circ T_2 \circ T_1(\mathbf{x}_j) \tag{1}$$

that is both short (meaning that the total number of transformations k is small), and well connected (meaning that the distance between two consecutive patterns along the path is also small). These sequences are referred to as consistent.

Assumption 2 *If two patterns $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different classes ($y_i \neq y_j$), all possible sequences of transformations between $\mathbf{x}_i$ and $\mathbf{x}_j$ are either very long (meaning that $k \gg 1$), or are not well connected (meaning that the connecting path contains at least one weak link formed by two distant patterns).*

As a result of Assumption 1, each pattern can be obtained as a sequence of connected image transformations from any other pattern of the same class (see Fig. 1), and all patterns that compose a consistent sequence belong to the same class. These transformations should have limited flexibility, otherwise any two patterns could be transformed one into each other by a single transformation. Underlying Assumption 1 is the idea that all patterns of the same class lie on a low-dimensional manifold. On the other hand, Assumption 2 is closely related to the idea that the classes form clusters that are separated by zones of low density. In other words, Assumptions 1 and 2 are somewhat equivalent to the standard low-dimensional manifold and cluster assumptions that are usually considered in semi-supervised settings. However, as we will see, the idea of *connected transformations* brings a new perspective on how these assumptions should be exploited in an efficient way, especially when only a few labeled data are available.

3. PROPOSED CLASSIFIER

According to the model of connected transformations given by Eq. (1), a *global* path-based distance measure should be computed by taking into account the whole sequence of *local* deformations or dissimilarities, starting from an unlabeled pattern and reaching a labeled one. Therefore, the proposed classifier requires defining three stages or blocks: i) a local pairwise dissimilarity metric d_s, ii) a global distance d_c that measures the length of the path through all connected image transformations, and iii) a final classification step based on the proposed global distance.

In this paper we mainly focus on the first two stages that provide us with a robust density-based metric for semi-supervised classification. In particular, once a suitable distance has been computed, any of the nearest neighbor-based techniques can be used for classification. For simplicity we choose the 1-NN classifier that selects the class of the closest labeled example $\mathbf{x}_{nn} \in \mathcal{X}_l$, which is found as

$$\mathbf{x}_{nn} = \arg\min_{\mathbf{x}_i \in \mathcal{X}_l} d_c(\mathbf{x}_j, \mathbf{x}_i). \tag{2}$$

In the following we first review the image deformation model of [1] which is used to compute the (local) similarity between image pairs, and then the (global) ρ-connectivity distance, as proposed in [2], which allows to allows to find optimal paths of transformations.

Algorithm 1 Image deformation model (IDM) dissimilarity [1].

input: images $\mathbf{a}$ and $\mathbf{b}$.
initialize: $d = 0$
for each $p = 1, 2, \ldots, P$; $q = 1, 2, \ldots, Q$ **do**
$d = d + \min_{\substack{r \in \{1,\ldots,P\} \bigcap \{p-w,\ldots,p+w\} \\ s \in \{1,\ldots,Q\} \bigcap \{q-w,\ldots,q+w\}}} \|\mathbf{a}_{pq} - \mathbf{b}_{rs}\|^2$
end for
output: $d_{idm}(\mathbf{a}, \mathbf{b}) = d$.

3.1. Dissimilarity based on local deformations

We are interested in a dissimilarity measure that allows to compensate for small geometric intra-class deformations while retaining the larger inter-class differences. The literature on image deformation models is vast, from elastic matching techniques [11] to shape contour models [12]. Apart from being flexible enough (but not too flexible), the chosen transformation model should be computationally efficient. As a good tradeoff between all these requirements, we use the image distortion model (IDM) proposed by Keysers et al. in [1]. This model has a very simple implementation and it has been applied successfully in supervised handwritten digit recognition.

Let us denote two images taken from the complete data set as $\mathbf{a} = \{\mathbf{a}_{pq}\}$ and $\mathbf{b} = \{\mathbf{b}_{pq}\}$. The pixel positions are indexed by $p = 1, \ldots, P$ and $q = 1, \ldots, Q$. In case the images differ in size, a scaling is taken into account (see [1] for details). In general, $\mathbf{a}_{pq}, \mathbf{b}_{pq} \in \mathcal{R}^h$ are *hyperpixels* that can represent grey values ($h = 1$), color values ($h = 3$), the horizontal and vertical local image gradients ($h = 2$), or a larger pixel context (for instance, $h = 18$ if the 3×3 pixel contexts of the local gradients are considered).

A summary of the IDM computation is provided in Algorithm 1. Specifically, for each hyperpixel $\mathbf{a}_{pq}$ of image $\mathbf{a}$, IDM aims to find the optimally corresponding hyperpixel $\mathbf{b}_{rs}$ of image $\mathbf{b}$ in a neighborhood limited by a warp range w. The IDM dissimilarity is then calculated as the conventional Euclidean distance between the image $\mathbf{a}$ and the transformed image $\mathbf{b}$. Since the optimal warping of one hyperpixel does not affect the optimal warping of its neighboring hyperpixels, IDM is referred to as a zero-order model. Models of first and second order take into account one or two levels of neighboring pixels, which guarantees a smoother warping, at the cost of a much higher computational burden.

An important observation on IDM is that it is not a symmetric proximity measure, and therefore it cannot be used directly as a distance function. Furthermore, as preliminary experiments on the MNIST data set pointed out, the differences between $d_{idm}(\mathbf{a}, \mathbf{b})$ and $d_{idm}(\mathbf{b}, \mathbf{a})$ can be significant for certain pairs of examples, indicating that IDM is a less reliable dissimilarity measure for some patterns. Motivated by these observations, the distance measure used as a local dissimilarity in the proposed method consists of a "worst case" symmetric IDM, specifically

$$d_s(\mathbf{x}_i, \mathbf{x}_j) = \max\left(d_{idm}(\mathbf{x}_i, \mathbf{x}_j), d_{idm}(\mathbf{x}_j, \mathbf{x}_i)\right). \tag{3}$$

Algorithm 2 Connected image transformations (CIT) classifier.

input: data set $\mathcal{X}_l$ with labels $\mathcal{Y}_l$, and data set $\mathcal{X}_u$.
Construct $\mathcal{X} = \mathcal{X}_l \bigcup \mathcal{X}_u$.
Calculate hyperpixels of all $\mathbf{x}_i \in \mathcal{X}$.
for each $\mathbf{x}_i \in \mathcal{X}$, $\mathbf{x}_j \in \mathcal{X}_u$, **do**
  Calculate IDM dissimilarity $d_s(\mathbf{x}_i, \mathbf{x}_j)$ with (3) and Alg. (1).
end for
for each $\mathbf{x}_j \in \mathcal{X}_u$, **do**
  Compute closest labeled pattern $\mathbf{x}_{nn} \in \mathcal{X}_l$ with (4) and (2).
  Assign label y_{nn} of closest labeled pattern $\mathbf{x}_{nn}$ to $\mathbf{x}_j$.
end for
output: Estimated labels $\hat{\mathcal{Y}}_u$ corresponding to $\mathcal{X}_u$.

3.2. The ρ-connectivity global distance measure

In order to find optimal paths of transformations, we define $p_{i,j}$ to be a path of length $r = |p_{i,j}|$ that connects $\mathbf{x}_i$ with $\mathbf{x}_j$ through an arbitrary sequence of intermediate patterns: $p_{i,j} = \{\mathbf{x}_i, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_j\}$. All patterns composing a given path are in $\mathcal{X}_l \bigcup \mathcal{X}_u$ without distinguishing between labeled and unlabeled patterns. We use the notation $p_{i,j}(k)$ to denote the k-th pattern in the path, therefore $p_{i,j}(1) = \mathbf{x}_i$ and $p_{i,j}(r) = \mathbf{x}_j$.

Let $P_{i,j}$ denote the set of all paths starting at $\mathbf{x}_i$ and ending at $\mathbf{x}_j$. The original connectivity distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, proposed in [13] as a mean to improve clustering algorithms, was defined as the length of the longest link of the shortest path joining the two points. In this way, points that can be connected through high-density regions are close to each other. In order to increase noise robustness, this measure was subsequently modified in [2] to accumulate the weighted dissimilarities between points along the shortest path. This modified connectivity distance, denoted here as "ρ-connectivity", is defined as

$$d_c(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\rho} \ln \left(1 + \min_{p_{i,j} \in P_{i,j}} \sum_{k=1}^{|p_{i,j}|-1} \left(e^{\rho d_s(\mathbf{x}_k, \mathbf{x}_{k+1})} - 1 \right) \right), \tag{4}$$

where $d_s(\mathbf{x}_k, \mathbf{x}_{k+1})$ measures the dissimilarity between patterns $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$. In words, (4) estimates the "optimal" path joining $\mathbf{x}_i$ and $\mathbf{x}_j$ as a weighted sum of the dissimilarities between pairs of consecutive points along such a path. The parameter ρ controls how these small deformations must be weighted along the path: for $\rho = 0$ we simply obtain the shortest path along all deformations without any weighting, while for $\rho \to \infty$ only the worst link (i.e., the largest deformation) produced along the path is considered. A proper selection of ρ yields a trade-off between these two extremes. Specifically, a value of $\rho < \infty$ allows to denoise the metric at a global scale and increases its robustness against outliers and bridge points. This is the principal parameter of the proposed classifier and, in general, the optimal value of ρ is problem-dependent.

3.3. Algorithm summary and implementation

An overview of the proposed CIT algorithm is given in Algorithm 2. In terms of computational cost, the bottleneck operation of the proposed method is the pattern matching procedure of IDM. Taking into account that the local dissimilarity is never computed between pairs of labeled data, the total amount of data pairs in this step is $n^2 + 2mn$. Each of these calculations requires to search a grid of $(2w+1)^2$ neighbors for each of the $p \cdot q$ hyperpixels of the reference image. Since $m \ll n$, the computational complexity of this step can be approximated as $O(n^2 \cdot pq \cdot w^2 h)$.

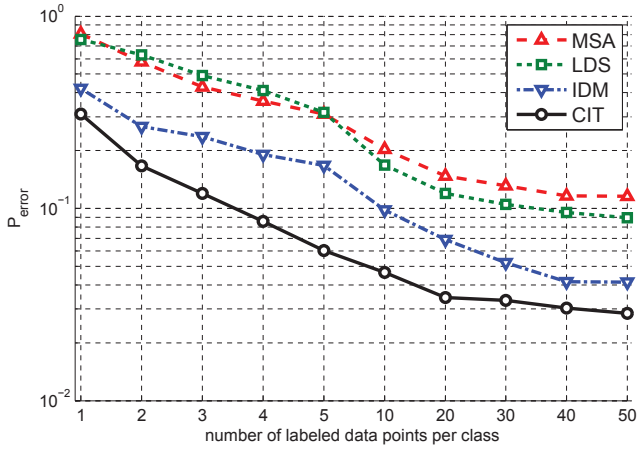

Fig. 2. Results on the MNIST data set using different numbers of labeled patterns and 1000 unlabeled patterns.

4. EXPERIMENTAL RESULTS

In this section we compare the results of four classification algorithms on the MNIST database. The MNIST database is the standard benchmark for handwritten character recognition[1]. All images contained in this database are gray-valued, sized 28×28, and they have been preprocessed by normalization and centering. The database contains a very large number of training and test data from 10 digit classes, 60000 and 10000 data points, respectively.

State-of-the-art supervised classification techniques obtain classification errors well below 1% when trained on all available MNIST training data. In particular, the k-NN classifier using IDM dissimilarities [1] obtains 0.54% error rate. In order to emphasize the capability of *semi-supervised* algorithms to exploit the information contained in the unlabeled data, we only use a very low number of labeled data in these experiments, together with a significant number of unlabeled data. The applied algorithms are the following:

1. MSA: The "Manifold Structure Approximation" technique from [9] aims to approximate each class as a low-dimensional manifold. This is achieved by calculating a low-rank approximation of the k nearest neighbor adjacency graph, whose weights are 1 between nearest neighbors and 0 otherwise.
2. LDS: The "Low Density Separation" algorithm from [2] introduces the ρ-connectivity distance to detect clusters that are separated by regions of low density, using the Euclidean distance as a local dissimilarity measure. It further exploits the unlabeled data by applying a transductive SVM classifier.
3. IDM: In [1] the "Image Deformation Model" dissimilarity measure was applied to construct a 1-NN classifier. We include its results to demonstrate that this supervised classifier is able to outperform some semi-supervised classifiers when only very few labeled patterns are available.
4. CIT: The proposed "Connected Image Transformations" classifier as described in Algorithm 2.

In Fig. 2 the classification results of the four algorithms are shown for varying numbers of labeled data patterns per class. These patterns were chosen randomly from the available training data for each class. In addition, 1000 unlabeled data patterns were used that

[1] `http://yann.lecun.com/exdb/mnist`.

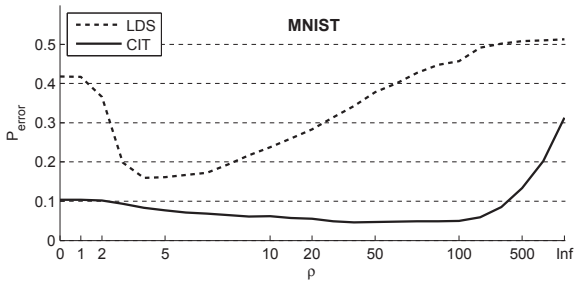

Fig. 3. Error probabilities obtained by applying the LDS and CIT classifiers with different values of ρ on the MNIST data set. For $\rho = 0$ the used distance is the shortest path length, while for $\rho = \infty$ the distance is the original connectivity distance from [13].

were chosen randomly from the entire unlabeled data pool. The parameters used for each algorithm are given in Table 1. The results were averaged out over 10 Monte-Carlo simulations. Using only $m = 10$ labeled data per class, CIT obtains an error rate of 4.64% here. To obtain the same error rate, IDM requires at least 3 times as many labeled patterns, while MSA and LDS do not even reach this rate in the tested range.

MSA	$k = 8$, $p = 0.2m$ (as in [9])
LDS	$\rho = 4$, rest as in [2]
IDM	$h = 18$, $w = 2$ (as in [1])
CIT	$h = 18$, $w = 2$, $\rho = 40$

Table 1. Parameters used for each algorithm.

As mentioned earlier, the principal parameter of the CIT classifier is ρ. This value depends on the particular data set and can be selected using cross-validation techniques, for instance. However, we have observed experimentally that optimal performance is typically achieved for a wide range of values. Fig. 3 shows the error rate versus ρ for the MNIST database, obtained for 10 labeled patterns per class and 1000 total unlabeled patterns. A similar behavior has been observed for other numbers of labeled and unlabeled patterns. In general, higher values of ρ assign more confidence to the weakest link in a path, while lower values can be used to average out several less reliable connections. This explains the difference between the optimal ρ for LDS, which uses an Euclidean local dissimilarity measure, and for CIT, which is based on the IDM measure. Finally, the paths plotted in Figures 4 and 5 allow to analyze the decisions taken by the proposed CIT classifier more closely.

5. CONCLUSIONS

We have proposed a semi-supervised handwritten digit classifier that is capable of operating with only very few labeled data available, using a distance measure that combines a local dissimilarity measure with a global connectivity-based metric. In experiments with subsets of the MNIST database, the presented classifier significantly outperforms other state-of-the-art semi-supervised algorithms.

Although the proposed classifier is easy to implement, it requires to calculate the IDM dissimilarity between all data points, which can be computationally costly. Therefore, future research topics include improving this local metric in terms of computational cost and accuracy, and also performing tests on other image databases.

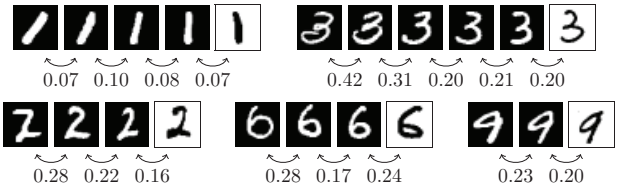

Fig. 4. Examples of paths followed for correctly classified digits. Each path starts at the unlabeled image to be classified, and after following several connections (other unlabeled images) it reaches the closest labeled image (depicted on a white background).

Fig. 5. Examples of paths followed for incorrectly classified digits.

6. REFERENCES

[1] D. Keysers, T. Deselaers, C. Gollan, and H. Ney, "Deformation models for image recognition," *IEEE Tr. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1422–1435, 2007.

[2] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *Proc. of the Tenth Intl. Workshop on Artificial Intelligence and Statistics*, 2005, pp. 57–64.

[3] J. Weston, C. Leslie, D. Zhou, A. Elisseeff, and W.S. Noble, "Semi-supervised protein classification using cluster kernels," in *Adv. in Neural Information Proc. Syst. (NIPS) 16*. 2004.

[4] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*, MIT Press, Cambridge, MA, 2006.

[5] Xiaojin Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[6] T.S. Jaakkola and S. Szummer, "Partially labeled classification with markov random walks," in *Advances in Neural Information Proc. Syst. (NIPS) 14*. MIT Press, 2002.

[7] A.D. Szlam, M. Maggioni, and R.R. Coifman, "Regularization on graphs with function-adapted diffusion processes," *Journal of Mach. Learning Research*, vol. 9, pp. 1711–1739, 2008.

[8] D. Zhou, O. Bousquet, T.M. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Proc. Syst. (NIPS) 16*. MIT Press, 2004.

[9] Belkin M. and Niyogi P., "Using manifold structure for partially labeled classification," in *Advances in Neural Information Proc. Syst. (NIPS) 15*, pp. 929–936. MIT Press, 2003.

[10] D. Francois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," *IEEE Trans. on Knowledge and Data Engineering*, vol. 19, no. 7, pp. 873 –886, jul. 2007.

[11] S. Uchida and H. Sakoe, "Eigen-deformations for elastic matching based handwritten recognition," *Pattern Recognition*, vol. 36, no. 9, pp. 2031–2040, 2003.

[12] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 24, pp. 509–522, 2002.

[13] B. Fischer, V. Roth, and J. M. Buhmann, "Clustering with the connectivity kernel," in *Advances in Neural Information Proc. Syst. (NIPS) 16*. MIT Press, 2004.