ISVC 2009

Distance Learning Based on Convex Clustering

Xingwei Yang[1], Longin Jan Latecki[1], and Ari Gross[2]

[1] Dept. of Computer and Information Sciences
Temple University, Philadelphia
{xingwei, latecki }@temple.edu
[2] Computer Science Dept.
Queens Collage, CUNY, New York
ari@cvisiontech.com

Abstract. Clustering has been among the most active research topics in machine learning and pattern recognition. Though recent approaches have delivered impressive results in a number of challenging clustering tasks, most of them did not solve two problems. First, most approaches need prior knowledge about the number of clusters which is not practical in many applications. Second, non-linear and elongated clusters cannot be clustered correctly. In this paper, a general framework is proposed to solve both problems by convex clustering based on learned distance. In the proposed framework, the data is transformed from elongated structures into compact ones by a novel distance learning algorithm. Then, a convex clustering algorithm is used to cluster the transformed data. Presented experimental results demonstrate successful solutions to both problems. In particular, the proposed approach is very suitable for superpixel generation, which are a common base for recent high level image segmentation algorithms.

1 Introduction

Clustering aims at finding hidden structure in a data set and is an important topic in machine learning and pattern recognition. Several methods, such as K-means [1], have been developed to solve the clustering problem for datasets which have a compact shape. However, they fail to handle data with complex cluster shapes, i.e., data that is not in the shape of point clouds, but instead forms curved and elongated shapes. Recently there have been some advances. Spectral Clustering [2] can handle this type of data very well and path-based clustering [3] also demonstrates excellent performance on some clustering tasks involving highly non-linear and elongated clusters in addition to compact clusters. However, these algorithms must have prior knowledge of the number of clusters, which is not practical in many applications.

In this paper, a new distance learning approach is proposed to transform elongated structures into compact ones. It is interleaved with a convex clustering method, which is used to find a global optimal solution for clustering. Apart from the Gaussian Kernel parameter, it is a completely parameter-free clustering principle. Learned distance are based on the convex clustering, which in turn are used for convex clustering, and so on. We have a natural stop criterion. We stop when cluster membership remains unchanged.

In addition to applying the proposed method to some toy examples, we demonstrate its excellent performance on image segmentation. The problem of segmenting an image into regions remains a great challenge for computer vision. Clustering has been crucial in attempting to solve this problem [3, 4]. We also show that the proposed approach is very suitable to over segment images to so called superpixels, where superpixels are small connected regions in the image. Commonly image segmentation algorithms like [4, 5] are used to generate superpixels. Recently many algorithms have been proposed for high level grouping of the superpixels, e.g., [6].

The rest of this paper is organized as follows. Some related work is briefly introduced in Section 2. In Section 3, the proposed distance learning and convex clustering approach will be introduced in detail. Experimental results will be given in Section 4 followed by the conclusion and discussion in Section 5.

2 Related work

There are several approaches to distance learning such as supervised distance metric learning and unsupervised distance metric learning. Most of the unsupervised distance metric learning methods are embedding methods, including linear and non-linear. The well known algorithms for nonlinear unsupervised dimensionality reduction are ISOMAP [7], Locally Linear Embedding(LLE) [8], and Laplacian Eigenamp (LE) [9]. ISOMAP seeks the subspace that best preserves the geodesic distances between any two data points, while LLE and LE focus on the preservation of local neighbor structure. Among the linear methods, Principal Component Analysis (PCA) [10] finds the subspace that best preserves the variance of data; Multidimensional Scaling (MDS) [11] finds the low-rank projection that best preserves the inter-point distance given by the pairwise distance matrix. In addition to the embedding methods, the path-based distance [3] transforms an original distance to a path based distance. It is defined as the min of the max over all paths. It can transforms the elongated data into compact data, but this approach relies on a greedy algorithm, which often leads to unsatisfactorily solutions. Outstanding results have been reported for the exemplar-based affinity propagation[12], which can automatically decide the number of clusters, but it cannot guarantee convergence, and therefore, it must be stopped manually.

Lashkari and Golland [13] have proposed a framework for constraining the search space of general mixture models to achieve global optimality of the solution, but this framework can deal only with data shaped as circular point clouds. We propose to first transform the data by learning new distance hierarchically so that the clusters in the new distance are more compact. This allows us to apply any classical clustering algorithm, yet for our applications we selected the algorithm by Lashkari and Golland, since it is guaranteed to yield optimal clustering and it automatically determines the number of clusters.

3 General framework for distance learning

Given a set of points $\mathcal{X} = \{x_1, \ldots, x_n\}$ and a distance function $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, where $\mathbb{R}_+$ is a set of nonnegative real numbers. We show that it is possible to learn a new distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ for any hierarchical clustering algorithm. Any kind of clustering method can be used to cluster the data points into k_l clusters $\{C_{l1}, \ldots, C_{lk_l}\}$ at level l, where $C_{0i} = \{x_i\}$ for $i = 1, \ldots, n$, and the number of clusters cannot increase, i.e., $k_l \geq k_{l+1}$. The distance at the beginning level, level 0, is the original distance among the input data points in $\mathcal{X}$. The distance between any data point a in C_{li} and any data point b in C_{lj} at level l for $i \neq j$ is:

$$d^{l+1}(a, b) = \min_{p \in C_{li}, q \in C_{lj}} D(p, q). \tag{1}$$

The distance $d^{l+1}(a, b)$ gives the learned distance of all data points in C_{li} to all data points in C_{lj}. If $a, b \in \mathcal{X}$ belong to the same cluster at level l, then $d^{l+1}(a, b) = d^l(a, b)$, i.e., the distances between points in the same cluster do not change. The hierarchical clustering process will automatically stop when the clusters in the level $l + 1$ are the same as in the level l. The new learned distance is then defined as

$$d(a, b) = d^l(a, b). \tag{2}$$

For brevity, we denote $d(x_i, x_j) = d_{ij}$ in the remainder of this paper.

Our intuition is that learned distance d can adequately represent the manifold structure of the data point set. Since the data points which have smaller distance will be clustered into one cluster instead of the two, it is consistent with the intuition that the data points should have small intra-cluster distance and large inter-cluster distance. Fig. 1 illustrates the intuition behind the hierarchical distance learning algorithm. It shows clustering results at each level of the approach together with the obtained new distances in the second row. If two points are in the same manifold, they will ultimately merge into one cluster which, and will have small distance.

We summarize the clustering based on the learned distance as: For a given input symmetric $(n \times n)$ matrix D of nonnegative pairwise dissimilarities between n objects, with zero diagonal elements, find clusters based on the learned distance d (instead of the original distances). Any clustering method could be used, but we use the clustering algorithm presented in the next section. In fact, we use this algorithm twice, first to learn new distances, and then to cluster the data based on the new distances.

4 From clustering to distance learning

We begin with an overview of the convex clustering approach in [13]. This approach can automatically determine the number of clusters for each level, and [13] have proved that their approach obtains a global optimal solution for a given Gaussian kernel. In each level of the convex clustering approach, q_{li} represents the probability that data point i in level l is the cluster center and

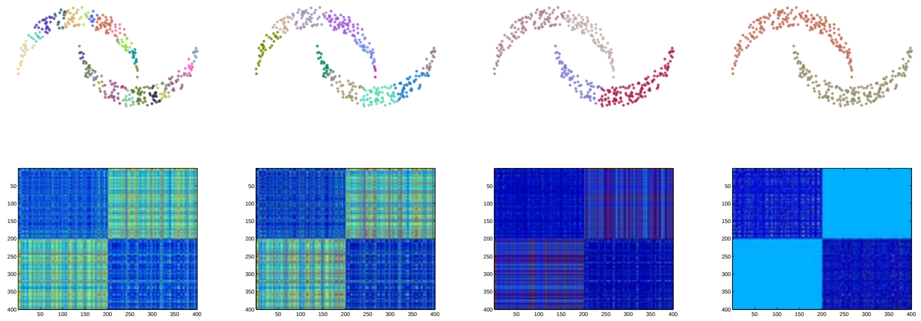


Fig. 1. Clustering results of levels 1, 2, 3, and 4. The corresponding learned distances are shown in the second row.

$s_{ij}^l = \exp(-\beta d_{ij}^l)$ represents the similarity between the two points i and j. According to the approach in [13], at each level the following two steps are iterated:

$$z_{li}^{(t)} = \sum_{j=1}^{n} s_{ij}^l q_{lj}^{(t)} \tag{3}$$

$$q_{lj}^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \frac{s_{ij}^l q_{lj}^{(t)}}{z_{li}^t} \tag{4}$$

From the view of point i, Eq. (3) represents how all the other points influence it. Eq. (4) represents, for a fixed point j, how much point j influences all the other points. In particular, for a pair of points j and i we have the ratio:

$$\frac{s_{ij}^l q_{lj}^{(t)}}{z_{li}^{(t)}} = \frac{s_{ij}^l q_{lj}^{(t)}}{\sum_{j=1}^{n} s_{ij}^l q_{lj}^{(t)}} \tag{5}$$

The term $q_{lj}^{(t)}$ represents the probability or strength of point j as a cluster center, the denominator $s_{ij}^l q_{lj}^{(t)}$ represents how strongly point i belongs to point j by the relation strength s_{ij}^l between them and, similarly, the numerator represents the total probability that point i belongs to all other points. Therefore, the ratio represents the normalized probability that point i belongs to point j. The higher the ratio, the more probable that point i belongs to point j.

As the procedure can find the optimal solution [13], the repetition will stop when the value $\sum_{j=1}^{n}(q_{lj}^{(t+1)} - q_{lj}^{(t)})$ is sufficiently small. Then, the soft assignment of point sets x_{li} to cluster center j, $r_{ij}^l = P(lj|x = x_{li})$ represents the probability distribution over the point set indices $lj : j = 1, \ldots, n$ and it is computed as

$$r_{ij}^l = \frac{s_{ij}^l q_{lj}^t}{z_{li}^{(t)}} \tag{6}$$

The point x_{li} is then assigned to a cluster based on

$$\text{assignment}(x_{li}) = \operatorname*{argmax}_j r^l_{ij} = \operatorname*{argmax}_j \frac{s^l_{ij} q^t_{lj}}{z^{(t)}_{li}} \tag{7}$$

The soft assignment in this paper is different from the method in [13]. In [13], all q_j that are below a certain threshold are set to zero and the entire distribution is renormalized over the remaining indices. This effectively excludes the corresponding points as possible exemplars and reduce the cost of the following iterations. However, in practice, the choice of threshold is very difficult and crucial for the results. For real applications, it is nearly impossible to choose a proper threshold. The proposed approach uses the hierarchical framework and therefore can use (7) directly to assign points to the cluster center, since the points which are incorrectly assigned will be corrected in the next level's clustering.

According to the assignment of (7), the algorithm automatically finds the number of clusters, which is represented as k_l. Based on (1), the distance between different clusters is then updated, and the clustering is repeated with the new distances. The process is repeated until new clusters are the same as clusters form the previous iteration.

5 Experimental results

In order to show the advantages of the framework of learning distance and the globally optimal clustering algorithm, two types of experiments were performed. The first type involves toy examples that illustrate the intuition behind our approach. The second type is the real-world application of image segmentation.

5.1 Toy examples

The results in Fig. 2 show the advantage of the proposed distance learning algorithm on clusters with more complex shapes than the two half moons in Fig. 1. There are 688 data points, each of which belong to one of three classes. Again the algorithm in [13], which works with the original Euclidean distances, is not able to yield correct clusters for any parameter settings. In order to show the effect of the proposed distance learning approach, Fig. 3 shows the distance matrix before and after the learning process, which represents all of the pairwise distances between points.

The above two toy examples show the features and the effectiveness of the proposed approach. In order to show that the proposed approach can be used in a real application, we implement the algorithm for image segmentation, which is still an unsolved problem in computer vision.

5.2 Image segmentation

Image segmentation is an integral part of image processing applications. Although in recent years image segmentation based on clustering algorithms have

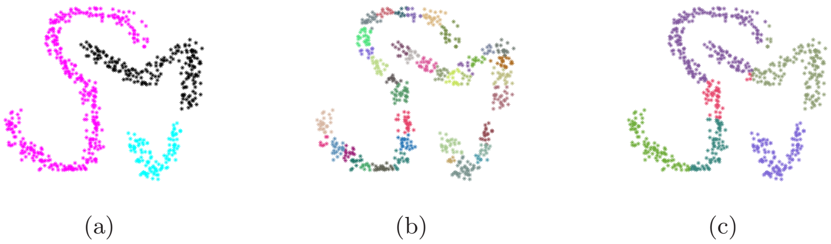
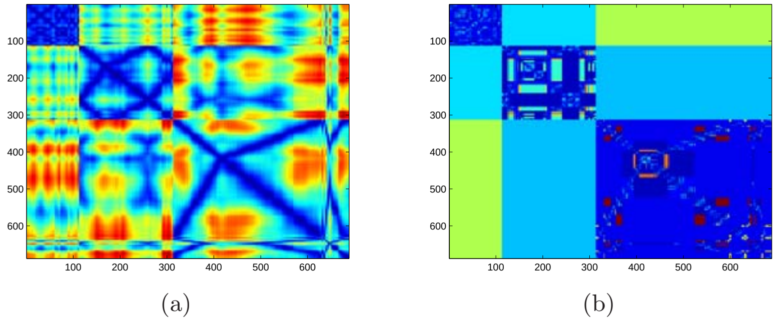

(a) (b) (c)

Fig. 2. (a) Clustering result of the proposed approach. (b) Clustering result of [13] with the same β as (a). (c) Clustering result of [13] with adjusted β.



(a) (b)

Fig. 3. (a) The distance matrix before the proposed distance learning approach. (b) The distance matrix after the proposed distance learning approach.

seen great success, the process is not fully automatic and the results are not as good as the vision community would like. However, by using the proposed distance learning algorithm and the modified clustering algorithm of [13], we can automatically find the segmentation results without a predefined number of clusters.

In our experiments, we use several images from the Weizemann dataset consisting of 328 horse images [14] of gray scale. Each pixel of an image is viewed as a data point and the distance between two pixels is the difference of their gray value and the Euclidean distance between their coordinates. Therefore, the s_{ij}^l in the above formulas can be obtained by $s_{ij}^l = \exp(-\beta_1 d_{1ij}^l) \cdot \exp(-\beta_2 d_{2ij}^l)$, where d_{1ij}^l is the difference of the gray value between two points at level l and d_{2ij}^l is the Euclidean distance between two points at level l. If the image's size is $m \times n$, the distance matrix will be $(m \cdot n)^2$, which is too high (reducing the storage requirement will be addressed in future work). Therefore, in our experiments the test image are relatively small. However, they still demonstrate the advantage of the proposed distance learning algorithm and the proposed clustering algorithm. Fig. 4 shows segmentation results of our method compared to [13] on our first test image. It is obvious that the proposed approach, as shown in Fig. 4(c) finds the optimal segmentation results automatically, whereas [13], as shown in Fig. 4(b), cannot find any informative segmentation results for this image.

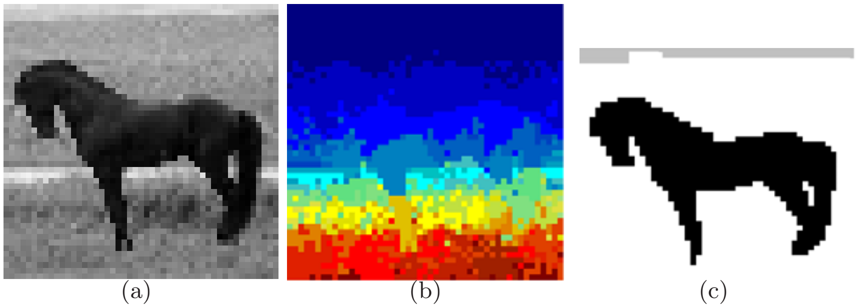


Fig. 4. (a) The original image. (b) Segmentation result of [13]. (c) Segmentation result of the proposed approach.

We also compared the image segmentation results of the proposed approach with the path-distance algorithm [15]. For both approaches, we first used the corresponding algorithms to obtain new distances, and then used our modified version of the clustering algorithm in [13] to segment the images. As can be seen in Fig. 5, the new distances learned by the proposed method perform significantly better than the path-distance distances of [15]. From the original image, it can be seen that the gray value of the foreground and background is not uniformly distributed which makes the clustering very difficult. However, the proposed approach can still find the optimal segmentation results, which shows the robustness of our approach.

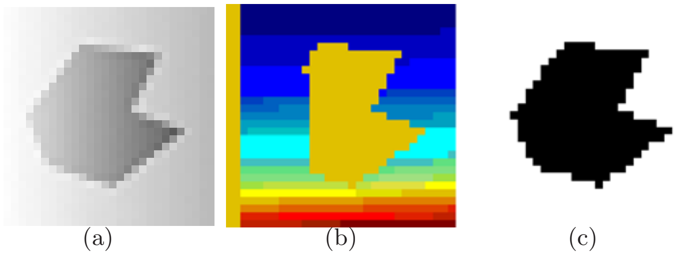


Fig. 5. (a) The original image from [16]. (b) Segmentation result based on distances learned according to [15]. (c) Segmentation result of the proposed approach.

Fig. 6 shows a few more examples of image segmentation with the proposed method on the Weizemann dataset [14]. The upper row shows the original image and the lower row shows the segmentation results obtained after 4,4,3 iterations, correspondingly. The morphological image postprocessing has been used to remove isolated segmentation pixels. The results are correct, even though only very elementary image information has been used: gray level differences and pixel distances. The middle row demonstrates the potential of the proposed method to generate superpixels, from left to right, we have 5, 4, 12 superpixels marked in different colors. These results are obtained after 3,3,2 iterations, correspondingly.

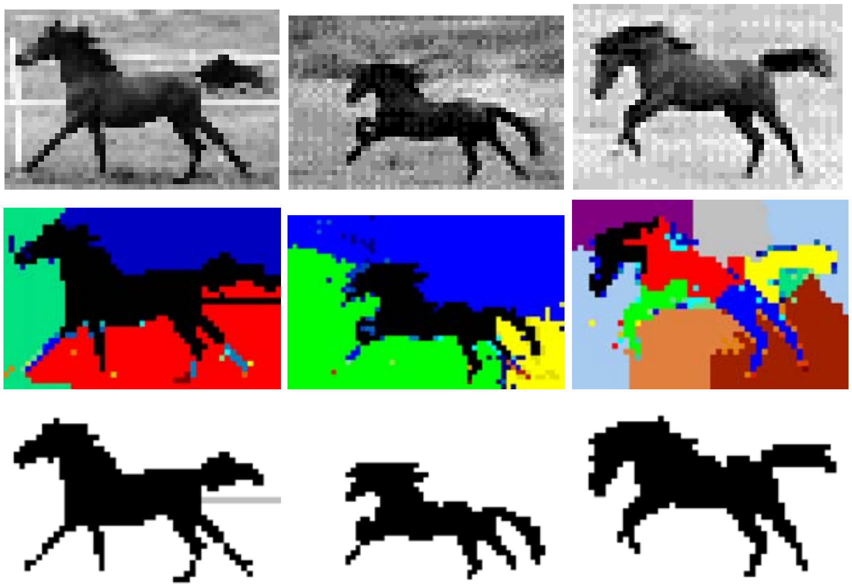
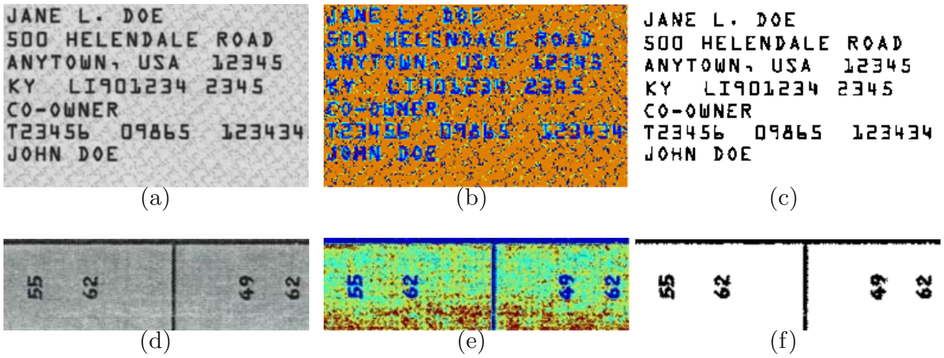


Fig. 6. Examples of image segmentation with the proposed method (bottom row) obtained after 4,4,3 iterations, correspondingly. We used only very elementary image information of gray level differences and pixel distances. The middle row show superpixels obtained by an intermediate stage of our algorithm after 3,3,2 iterations, correspondingly.

Besides the Weizemann dataset, Fig. 7 shows the performance on the document images. Fig. 7 (b) is the third iteration of the proposed approach, which could already distinguish the characters from the background, but it still contains many small components. The segmentation results of the final forth iteration is shown in (c). The image contains three clusters represented by black, white and gray colors. As the gray component is caused by the blur on the edge of character, it is always located on the edge of letters. Morphological postprocessing was used to remove isolated points. Similarly, Fig. 7 (e)-(f) shows the segmentation process for another document image (d).

The proposed approach can also be used in cell image segmentation. In Fig. 8, each cell images are segmented into 3 clusters. As the core of the cells are often different from other parts, it is reasonable for the algorithm to treat them differently. As for cell and document segmentation, the foreground objects are always separated and are sparse.

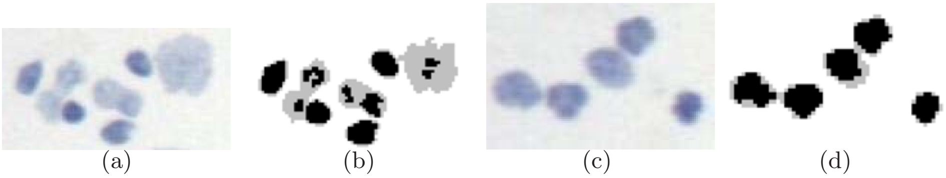
Though the proposed approach could work well on different kinds of images, the main drawback of the proposed approach is that the results are sensitive to the size of the Gaussian kernel. This is the trade-off for not having prior knowledge about the number of clusters.



(a) (b) (c)

(d) (e) (f)

Fig. 7. Segmentation results on document images (a,d). (b) and (e) are the intermediate steps of our segmentation. The final segmentation results are shown in (c) and (f).



(a) (b) (c) (d)

Fig. 8. (b) and (d) are the corresponding segmentation results of the cell images (a) and (c).

6 Conclusion

In this paper, we present a general framework to learn new distances by hierarchical clustering. At each level, a clustering algorithm is used to cluster data points. According to the cluster assignment, the distance is updated to extract the manifold structure of the data. A convex clustering approach is used as the algorithm for learning the distance. Based on the learned distance, the convex clustering algorithm can automatically find the globally optimal solution even for non-linearly distributed data. Our method does not involve any parameter except the parameter for Gaussian kernel. In particular, the number of clusters is determined automatically, which is an important advantage over most of the existing clustering algorithms.

References

1. Han, J., Kamber, M.: Data mining: Concepts and techniques. Morgan Kaufmann (2000)
2. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems. Volume 14. (2002) 849–856
3. Fischer, B., Buhmann, J.M.: Path-based clustering for grouping of smooth curves and texture segmentation. IEEE. PAMI **25** (2003) 513–518
4. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision **59** (2004) 167–181
5. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE. PAMI (2000)
6. Moore, A., Prince, S., Warrell, J., U.Mohammed, Jones, G.: Superpixel lattices. In: CVPR. (2008)
7. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science (2000) 290
8. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science (2000)
9. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation **15** (2003)
10. Gonzales, R., Woods, R.: Digital image processing. Addison-Wesley (1992)
11. Cox, T., Cox, M.: Multidimensional scaling. Chapman and Hall (1994)
12. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315** (2007) 972–976
13. Lashkari, D., Golland, P.: Convex clustering with exemplar-based models. In: Advances in Neural Information Processing Systems. (2007)
14. Borenstein, E., Shron, E., Ullman, S.: Combining top-down and bottom-up segmentation. In: Proc. IEEE workshop on Perc. Org. in Com. Vis. (2004)
15. Fischer, B., Roth, V., Buhmann, J.M.: Clustering with the connectivity kernel. In: Advances in Neural Information Processing Systems. (2004)
16. Chen, Q., Sun, Q., Heng, P.A., Xia, D.: A double-threshold image binarization method based on edge detector. Pattern reconition **41** (2008) 1254–1267