Hierarchical Semi-Supervised Clustering using KSC based model

Siamak Mehrkanoon, Oscar Mauricio Agudelo, Raghvendra Mall and Johan A. K. Suykens

Abstract— **This paper introduces a methodology to incorporate the label information in discovering the underlying clusters in a hierarchical setting using multi-class semi-supervised clustering algorithm. The method aims at revealing the relationship between clusters given few labels associated to some of the clusters. The problem is formulated as a regularized kernel spectral clustering algorithm in the primal-dual setting. The available labels are incorporated in different levels of hierarchy from top to bottom. As we advance towards the lowers levels in the tree all the previously added labels are used in the generation of the new levels of hierarchy. The model is trained on a subset of the data and then applied to the rest of the data in a learning framework. Thanks to the previously learned model, the out-of-sample extension property of the model allows then to predict the memberships of a new point. A combination of an internal clustering quality index and classification accuracy is used for model selection. Experiments are conducted on synthetic data and real image segmentation problems to show the applicability of the proposed approach.**

I. Introduction

In many applications, ranging from data mining to machine perception, labeled data is often sparse because it is both time consuming and costly to obtain. Therefore in many cases one encounters a large amount of unlabeled data while the labeled data are rare. So the first challenge is to be able to make use of labeled data points to boost the performance with respect to a purely unsupervised algorithm. In these cases one may consider the semi-supervised learning framework which concerns the problem of learning in the presence of both labeled and unlabeled data [1], [2], [3], [4].

The second challenge is how to associate the discovered clusters. In many application from biological data to web image organization, there would be a large number of clusters and sub-clusters. Some of sub-clusters may be relevant and of the others may belong to a more general cluster. In addition, some of the sub-clusters may or may not have labels. Therefore the incorporation of the available labels for presenting a relationship between clusters in a hierarchical fashion would be necessary.

Most of the developed semi-supervised approaches attempt to improve the performance by incorporating the information from either the unlabeled or labeled part. Among them are graph based methods that assume that neighboring point pairs with a large weight edge are most likely within the same cluster. The Laplacian support vector machine (LapSVM) [5], a state-of-the-art method in semi-supervised classification, is one of the graph based methods which provides a natural out-of-sample extension.

Some semi-supervised clustering methods have been focused on the use of side-information in the form of instance level must-link and cannot-link constraints. A must-link (ML) constraint enforces that two instances must be placed in the same cluster while a cannot-link (CL) constraint enforces that two instances must not be placed in the same cluster. However, both ML and CL constraints are not suitable for hierarchical clustering methods since objects are linked over different hierarchy levels [6], [7].

Spectral clustering methods belong to a family of unsupervised learning algorithms that make use of the eigenspectrum of the Laplacian matrix of the data to divide a dataset into natural groups such that points within the same group are similar and points in different groups are dissimilar to each other [8], [9], [10]. In this approach the eigenvectors become a new representation of the data, where the clusters form a localized structure. Finding the final grouping in the eigen-space is typically done by applying simple clustering techniques such as k-means.

Kernel spectral clustering (KSC) is an unsupervised algorithm introduced in [11]. The primal problem of the kernel spectral clustering is formulated as a weighted kernel PCA [12]. In contrast to classical spectral clustering, there is a systematic model selection scheme for tuning the parameters and also the extension of the clustering model to out-of-sample points is possible. KSC provides a partitional clustering, i.e. the dataset is decomposed into a number of disjoint clusters which are optimal in terms of some predefined internal quality index. However in some applications, a more informative hierarchical representation of the data is desirable. Hierarchical clustering groups the data points into a hierarchical tree-like structure using bottom-up or top-down approaches. The authors in [13] introduced Hierarchical Kernel Spectral Clustering where the Fisher criterion is used to reveal the hierarchical structure of the data.

A binary semi-supervised classification formulation is proposed in [14]. A non-parallel semi-supervised classifiers that generates two non-parallel hyperplanes by learning from both labeled and unlabeled data points is introduced in [15]. Recently Mehrkanoon *et al.* [16] proposed a multi-class semi-supervised algorithm (MSS-KSC) where KSC is used as a core model. MSS-KSC is a regularized version of KSC which aims at incorporating the information of the labeled data points in the learning process. The method can be applied for both semi-supervised classification and clustering and uses a low-dimensional embedding. In the MSS-KSC approach, one needs to solve a linear system of equations to obtain the model parameters. Therefore with n number of training points, the algorithm has $\mathcal{O}(n^3)$ training complexity

The authors are with the Department of Electrical Engineering ESAT-STADIUS, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium (email: {siamak.mehrkanoon}@esat.kuleuven.be).

with naive implementations. The MSS-KSC model can be trained on a subset of the data (training data points) and then applied to the rest of the data in a learning framework. Thanks to the previously learned model, the out-of-sample extension property of the MSS-KSC model allows then to predict the memberships of a new point. Moreover, as it has been shown in [17], it can scale to large data sets.

It is the purpose of this paper to develop a hierarchical clustering (HMSS-KSC) based on the semi-supervised KSC approach. We show how one can incorporate the available side-information (labels) in different hierarchy levels. The hierarchical representation is based on the internal cluster validation index which quantifies the goodness of the clustering. The cluster validation index can reveal several clustering model parameters leading to good clusterings. In the final stage, these clustering results are combined to display the underlying cluster hierarchies and relation among clusters found. This can play an important role in discovering multiscale structure present in the data.

This paper is organized as follows. In Section II, classical hierarchical clustering methods are briefly reviewed. In Section III, an overview of the multi-class semi-supervised clustering (MSS-KSC) algorithm is given. The hierarchical semi-supervised clustering HMSS-KSC is described in Section IV. In Section V, experimental results are given in order to confirm the validity and applicability of the proposed method.

II. Hierarchical clustering

In the classical hierarchical clustering, a dendrogram is obtained based on a dissimilarity measure between each pair of observations and most often the Euclidean distance is used. The algorithm proceeds iteratively by starting at the bottom of the dendrogram, where each of the observations is treated as one cluster. Then the two clusters that are most similar to each other are merged.

Pairs of clusters are then merged as the hierarchy goes up in the tree. Each merge is represented by a horizontal line and the y-axis indicates the similarity (or dissimilarity) of the two merging clusters. The algorithm proceeds in this fashion until all of the observations belong to one single cluster. The linkage measure, which defines the dissimilarity between two groups of observations, determines which clusters should fuse. The four most common types of linkage are complete, average, single and centroid.

Single linkage computes the minimal intercluster dissimilarity and is a local criterion taking into account only the zone where the two clusters are closest to each other. Complete linkage computes the maximal intercluster dissimilarity and is a non-local criterion giving preference to compact clusters. Both single and Complete have been reported to have some drawbacks. This Single criterion suffers from an undesirable effect called chaining. Chaining causes unwanted elongated clusters since the overall shape of the formed clusters is not taken into account. On the other hand the complete linkage also suffers from high sensitivity to outlying data points. Average linkage which computes the mean intercluster dissimilarity is a specialized method trying to find a compromise between single and complete linkage.

The authors in [13] proposed a methodology based on KSC to discover cluster hierarchies. The BLF [11] criterion is used to select the optimal model parameter pairs i.e. number of clusters k and kernel bandwidth σ. In this paper we will propose a new method for reveal the underlying hierarchical structure of the data given a few amount of labeled data points based on MSS-KSC briefly reviewed in Section III.

III. Semi-Supervised Clustering using MSS-KSC

Consider training data points

$$\mathcal{D} = \{\underbrace{x_1, ..., x_{n_{UL}}}_{\substack{Unlabeled \\ (\mathcal{D}_U)}}, \underbrace{x_{n_{UL}+1}, .., x_n}_{\substack{Labeled \\ (\mathcal{D}_L)}}\}, \tag{1}$$

where $\{x_i\}_{i=1}^n \in \mathbb{R}^d$. The first n_{UL} data points do not have labels whereas the last $n_L = n - n_{UL}$ points have been labeled. Assume that there are Q classes, then the label indicator matrix $Y \in \mathbb{R}^{n_L \times Q}$ is defined as follows:

$$Y_{ij} = \begin{cases} +1 & \text{if the } i\text{th point belongs to the } j\text{th class} \\ -1 & \text{otherwise.} \end{cases} \tag{2}$$

The Multi-class semi-supervised KSC (MSS-KSC) described in [16] is formulated as follows:

$$\begin{aligned} \min_{w^{(\ell)}, b^{(\ell)}, e^{(\ell)}} \quad & \frac{1}{2}\sum_{\ell=1}^{Q} w^{(\ell)^T} w^{(\ell)} - \frac{\gamma_1}{2}\sum_{\ell=1}^{Q} e^{(\ell)^T} V e^{(\ell)} + \\ & \frac{\gamma_2}{2}\sum_{\ell=1}^{Q} (e^{(\ell)} - c^{(\ell)})^T \tilde{A} (e^{(\ell)} - c^{(\ell)}) \\ \text{subject to} \quad & e^{(\ell)} = \Phi w^{(\ell)} + b^{(\ell)} 1_n, \ \ell = 1, \ldots, Q, \end{aligned} \tag{3}$$

where c^ℓ is the ℓ-th column of the matrix C defined as

$$C = [c^{(1)}, \ldots, c^{(Q)}]_{n \times Q} = \left[\frac{0_{n_{UL} \times Q}}{Y} \right]_{n \times Q}, \tag{4}$$

where $0_{n_{UL} \times Q}$ is a zero matrix of size $n_{UL} \times Q$ and Y is defined as previously. The matrix $\tilde{A}$ is defined as follows:

$$\tilde{A} = \left[\begin{array}{c|c} 0_{n_{UL} \times n_{UL}} & 0_{n_{UL} \times n_L} \\ \hline 0_{n_L \times n_{UL}} & I_{n_L \times n_L} \end{array} \right],$$

where $I_{n_L \times n_L}$ is the identity matrix of size $n_L \times n_L$. V is the inverse of the degree matrix defined as previously.

Since in (3) the feature map φ is in general assumed to be not explicitly known, one uses the kernel trick and solves the problem in the dual. The Lagrangian of the constrained optimization problem (3) becomes

$$\mathcal{L}(w^{(\ell)}, b^{(\ell)}, e^{(\ell)}, \alpha^{(\ell)}) = \frac{1}{2}\sum_{\ell=1}^{Q} w^{(\ell)^T} w^{(\ell)} - \frac{\gamma_1}{2}\sum_{\ell=1}^{Q} e^{(\ell)^T} V e^{(\ell)}$$

$$+ \frac{\gamma_2}{2}\sum_{\ell=1}^{Q} (e^{(\ell)} - c^{(\ell)})^T \tilde{A} (e^{(\ell)} - c^{(\ell)}) +$$

$$\sum_{\ell=1}^{Q} \alpha^{(\ell)^T} \left(e^{(\ell)} - \Phi w^{(\ell)} - b^{(\ell)} 1_n \right),$$

where $\alpha^{(\ell)}$ is the vector of Lagrange multipliers. Then the Karush-Kuhn-Tucker (KKT) optimality conditions are as follows,

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w^{(\ell)}} = 0 \rightarrow \; w^{(\ell)} = \Phi^T \alpha^{(\ell)}, \; \ell = 1, \ldots, Q, \\ \frac{\partial \mathcal{L}}{\partial b^{(\ell)}} = 0 \rightarrow \; 1_n^T \alpha^{(\ell)} = 0, \; \ell = 1, \ldots, Q, \\ \frac{\partial \mathcal{L}}{\partial e^{(\ell)}} = 0 \rightarrow \; \alpha^{(\ell)} = (\gamma_1 V - \gamma_2 \tilde{A}) e^{(\ell)} + \gamma_2 c^{(\ell)}, \ell = 1, \ldots, Q, \\ \frac{\partial \mathcal{L}}{\partial \alpha^{(\ell)}} = 0 \rightarrow \; e^{(\ell)} = \Phi w^{(\ell)} + b^{(\ell)} 1_n, \; \ell = 1, \ldots, Q. \end{cases} \tag{5}$$

Elimination of the primal variables $w^{(\ell)}, e^{(\ell)}$ and making use of Mercer's Theorem [18], results in the following linear system in the dual [16]:

$$\gamma_2 \left(I_n - \frac{R 1_n 1_n^T}{1_n^T R 1_n} \right) c^{(\ell)} = \alpha^{(\ell)} - R \left(I_n - \frac{1_n 1_n^T R}{1_n^T R 1_n} \right) \Omega \alpha^{(\ell)}, \tag{6}$$

where $R = \gamma_1 V - \gamma_2 \tilde{A}$. As is is shown in [16], given Q labels the approach is not restricted to finding just Q classes and instead is able to discover up to 2^Q hidden clusters. In addition, it uses low embedding dimension to reveal the existing number of clusters which is important when one deals with large number of clusters. In fact one maps the data points to a Q-dimensional space, which from now on will be referred to as α-space, and the solution vectors $\alpha^{(\ell)}$ ($\ell = 1, \cdots, Q$) represent the embedding of the input data in this space. Therefore every point x_i is associated with the point $[\alpha_i^{(1)}, \cdots, \alpha_i^{(Q)}]$ in the α-space. (space spanned by the solution vector $\alpha^{(\ell)}$). The MSS-KSC algorithm [16] is summarized in Algorithm 1.

IV. Hierarchical Semi-supervised clustering

A. Methodology

The kernel spectral clustering (KSC) algorithm proposed in [11] is provided with a model selection procedure based on the Balanced Line Fit (BLF) criterion. It can be shown that in the ideal situation of well separated clusters, the data projections (score variables e_i) associated with the KSC formulation, form lines one per each cluster. The shape of the data points in the projections space, is exploited by the BLF criterion to select the optimal clustering parameters e.g. the number of clusters (k) and the kernel bandwidth σ. The BLF criterion is defined as follows:

$$\text{BLF}(\mathcal{D}^{\text{Val}}, k) = \eta \text{linefit}(\mathcal{D}^{\text{Val}}, k) + (1 - \eta) \text{balance}(\mathcal{D}^{\text{Val}}, k) \tag{7}$$

where $\mathcal{D}^{\text{Val}}$ represents the validation set and k indicates the number of clusters. The linefit index equals 0 when the score variables are distributed spherically and equals 1 when the score variables are collinear, representing points in the same cluster. The balance index equals 1 when the clusters have the same number of elements and tends to 0 in extremely unbalanced cases. The parameter η controls the importance given to the linefit with respect to the balance index and takes values in the range $[0, 1]$.

Algorithm 1: MSS-KSC approach [16]

```
Input:  Training data set $\mathcal{D}$, labels $Y$, the tuning
        parameters $\{\gamma_i\}_{i=1}^2$, the kernel parameter (if
        any), number of clusters $k$, the validation set
        $\mathcal{D}^{val} = \{x_i^{val}\}_{i=1}^{N_{val}}$ and number of available
        class labels i.e. $Q$
Output: Cluster membership of validation data points
        $\mathcal{D}^{val}$

Solve the dual linear system (6) to obtain $\{\alpha^\ell\}_{\ell=1}^Q$ and
  compute the bias term $\{b^\ell\}_{\ell=1}^Q$.
Binarize the solution matrix
  $S_\alpha = [\text{sign}(\alpha^{(1)}), \ldots, \text{sign}(\alpha^{(Q)})]_{M \times Q}$, where
  $\alpha^\ell = [\alpha_1^\ell, \ldots, \alpha_M^\ell]^T$.
Form the codebook $\mathcal{CB} = \{c_q\}_{q=1}^p$, where
  $c_q \in \{-1, 1\}^Q$, using the $k$ most frequently occurring
  encodings from unique rows of solution matrix $S_\alpha$.
Estimate the score variables for the validation data
  points $\{e_{val}^{(\ell)}\}_{\ell=1}^Q$.
Binarize the validation projections and form the
  encoding matrix $[\text{sign}(e_{val}^{(1)}), \ldots, \text{sign}(e_{val}^{(Q)})]_{N_{val} \times Q}$ for
  the test points (Here $e_{val}^\ell = [e_{val,1}^\ell, \ldots, e_{val,N_{val}}^\ell]^T$).
$\forall i$, assign $x_i^{val}$ to class/cluster $q^*$, where
  $q^* = \underset{q}{\text{argmin}} \, d_H(e_{val,i}^\ell, c_q)$ and $d_H(\cdot, \cdot)$ is the Hamming
  distance.
```

It was shown in [13] that the BLF criterion has multiple peaks corresponding to different values of kernel parameter σ for given number of clusters k. For the semi-supervised setting, given the fact that Q labels are available and we can detect maximum up to 2^Q clusters (see [16]), the criterion introduced in [16] is employed on the cluster intervals $[2^{i-1}+1, 2^i]$ for $i = 2, \ldots, Q$ to detect the ideal number of clusters k for each level of hierarchy in the given dataset.

Therefore a grid search over different values of $k \in [2^{i-1}+1, 2^i]$ and σ evaluating the internal Cluster Validation Index (CVI) on validation data is performed. The parameter pairs (k, σ) for which the CVI is larger than a specified threshold value θ are selected. Here θ is set to three fourths of a maximum value of CVI. We then build the MSS-KSC model using the optimal parameter pairs (k,σ) and obtain the cluster memberships for all the points using the out-of-sample extensions property. The flow-chart of the proposed algorithm is shown in Fig. 1. In constructing the hierarchy we start with a large number of values of k and afterwards moving to intervals with smaller value of k. Thus the hierarchy of clusters are obtained in a bottom-up fashion. The proposed methodology is outlined in Algorithm 2. In Algorithm 2, a small portion of the data points is used as training set for learning the model. Considering that the total number of data points is N and we only work with n training points, $n \ll N$, the computational complexity of the HMSS-KSC algorithm is $O(n^3 + Nn)$. In agglomerative hierarchical clustering algorithms like linkage techniques, one starts with the whole dataset and therefore at the lowest level of hierarchy the complexity for obtaining the pairwise similarities becomes $O(N^2)$.

Algorithm 2: Hierarchical MSS-KSC

```
Input: Training data set $\mathcal{D}$, labels $Y$, set of $r$ values
       for the kernel parameter $\sigma \in \{\sigma_1, \ldots, \sigma_r\}$,
       number of available class labels i.e. $Q$
Output: Cluster membership of validation data points
       $\mathcal{D}^{val}$
for $i \leftarrow 1$ to $Q$ do
    cluster-range=$[2^{i-1}+1, 2^i]$
    for $k \in$ cluster-range do
        $\forall$ combination of parameter pairs $(k, \sigma)$ where
        $\sigma \in [\sigma_1, \ldots, \sigma_r]$ train the MSS-KSC algorithm
        $1$ with $i$ labels.
        Find the maximum value of the cluster quality
        index (CQI) over the range $\sigma$ values.
        If the maximum value of CQI is greater than the
        threshold $\theta$, select the optimal parameter pairs
        $(k^*, \sigma^*)$ and build a level of cluster hierarchy
        Use the training data $\mathcal{D}$ and $(k^*, \sigma^*)$ to train the
        MSS-KSC in algorithm 1 and compute the
        out-of-sample extension for the test points.
return The cluster memberships for all the selected
       levels of hierarchy.
```

B. Hierarchical semi-supervised representation

After obtaining the clusters for each level of hierarchy, a specialized linkage criterion similar to the that described in [13] determines which clusters are merging based on the evolution of the cluster memberships as the hierarchy goes up. The y-axis of the dendrogram represents the scale and corresponds to the value of the kernel parameter σ at which the merge occurs. During the merging, some data points of the merging clusters might go to a non-merging cluster. The remaining data points are then forced to join the merging cluster of the majority. Each leaf of the dendrogram represents one cluster. However, as we move up the tree, some leaves begin to fuse into branches. These correspond to clusters that are similar to each other. As the number of clusters k increases, more complex models are needed to discover the structure of the data, therefore usually the optimal kernel bandwidth σ decreases. Usually the clusters that appears lower in the dendrogram are more complex and clusters that fuse later (near the top of the tree) are less complex.

The available side-information (labeled data points) is improving the clustering results. In order to show how the labels are incorporated into the results of a hierarchy, for levels of the resulting tree that labels are used, we also plot the labels to indicate the presence of the labels as we go down the tree. As the MSS-KSC algorithm is able to find up to 2^Q clusters when Q labels is used, some of the levels in the tree will not have labels but still the algorithm is able to discover the hidden clusters present in the dataset.

V. Numerical Experiments

In this section, some experimental results are presented to illustrate the applicability of the proposed hierarchical MSS-KSC algorithm. In order to illustrate the effect of prototypes (labels), we start with two synthetic problems and show how the labels can be incorporated to the learning process and affects the clustering results. Next we show the experimental results on some color images from the Berkeley image data set [1]. For the Berkeley images data set for which the ground truth segmentations are known, the segmentations obtained by HMSS-KSC and are compared with the ground truth in Table II. The following semi-supervised model selection criterion.

$$\underset{\gamma_1,\gamma_2,\sigma}{\text{argmax}}\ \kappa\, \mathit{CLP}(\gamma_1,\gamma_2,\sigma) + (1-\kappa)\mathit{Acc}(\gamma_1,\gamma_2,\sigma) \qquad (8)$$

where *CLP* and *Acc* stand for clustering performance and classification accuracy respectively. $\kappa \in [0, 1]$ is a user-defined parameter that controls the trade-off between the importance given to unlabeled and labeled samples. A common approach for evaluation of clustering results is to use cluster validity indices [19], [9], [8]. Any internal clustering validity approach such as Silhouette index [20], Davies-Bouldin index (DB) or BLF [11] can be utilized. In this paper the Silhouette index is used.

A. Synthetic data sets

We start with a synthetic data set consisting of six well separated Gaussians. Some labeled data points from three of them are available (see Fig. 2). The labels are feed to the algorithm at different levels of the tree. Fig. 2(g) shows the order of presenting the labels. First the information of the labeled point shown by □ is used at the top most level. In the next level the data point denoted by △ is provided which enables the algorithm to discover up to 4 clusters. The selection of number of clusters is performed which indicates that three clusters is a good candidate (according to the model selection criterion (8). Next the last label $*$ is used and therefore the blue cluster splits in two. At this point that all three labels are presented to the algorithm, ideally up to eight clusters can be found. Hence the other hidden clusters that have no labels are also detected. From Fig. 2(g), one may notice that, two splits happened at the same level (using the same kernel bandwidth σ). Although this is no longer a dendrogram but a tree structure, one still can think of the splitting order (merge order). This pattern depends on the structure of the data under study and the more nonlinear the clusters, the less likely that it occurs.

In Fig. 3, there is a happy face which can be clustered into five groups based on their colors. We provide three labels from three clusters (the background, the face and the eyes). The first label from the background is used at the top level of tree to partition the image into two groups. In the next

[1] Available at: http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

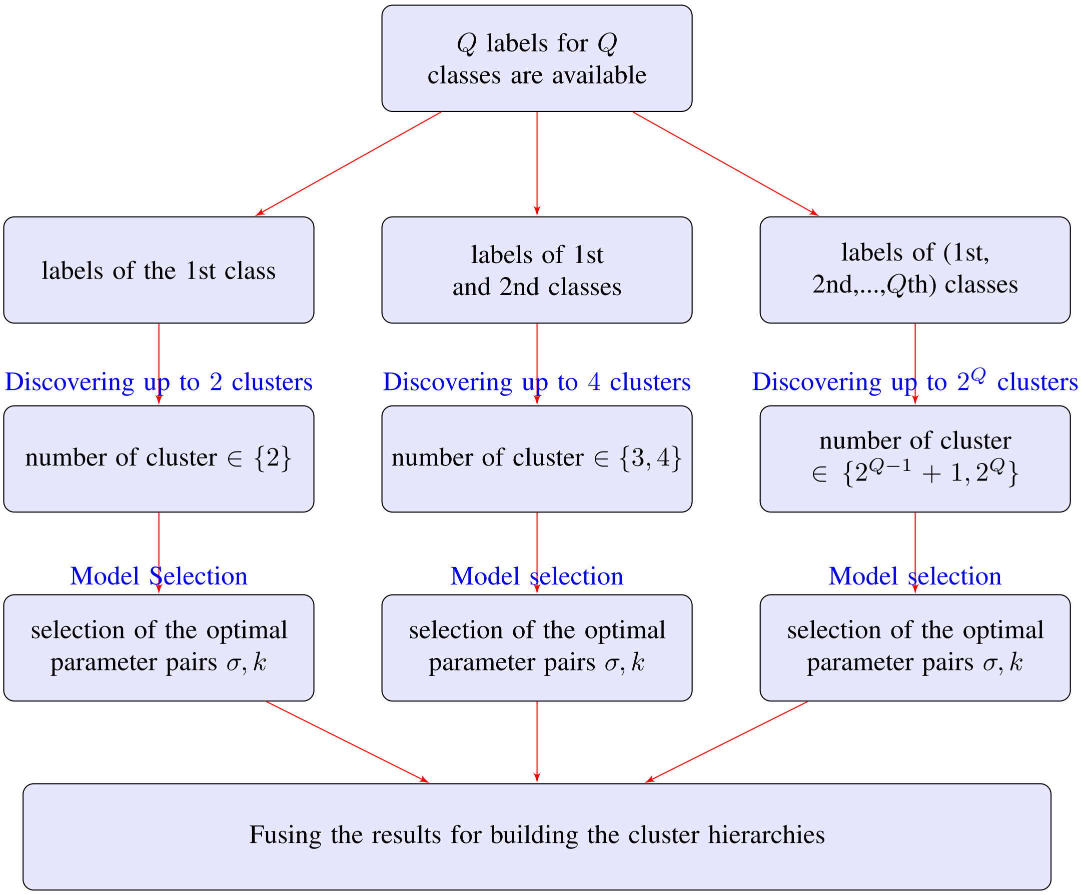

Q labels for Q classes are available
labels of the 1st class
labels of 1st and 2nd classes
labels of (1st, 2nd,...,Qth) classes
Discovering up to 2 clusters
Discovering up to 4 clusters
Discovering up to 2^Q clusters
number of cluster $\in \{2\}$
number of cluster $\in \{3, 4\}$
number of cluster $\in \{2^{Q-1}+1, 2^Q\}$
Model Selection
Model selection
Model selection
selection of the optimal parameter pairs σ, k
selection of the optimal parameter pairs σ, k
selection of the optimal parameter pairs σ, k
Fusing the results for building the cluster hierarchies

Fig. 1. Flow-chart of the Hierarchical Multi-class Semi-Supervised Kernel Spectral Clustering (HMSS-KSC) algorithm.

level the label of the face is added. Therefore totally there will be two labels present for the algorithm and thus up to four clusters can potentially be detected. Next we feed the last label, which indicates one of the eyes of the face, to the algorithm. Given that three labels are now provided up to eight clusters can be detected. The optimal model parameters, number of clusters k and kernel bandwidth σ are obtained based on the (8).

B. Hierarchical Image segmentation

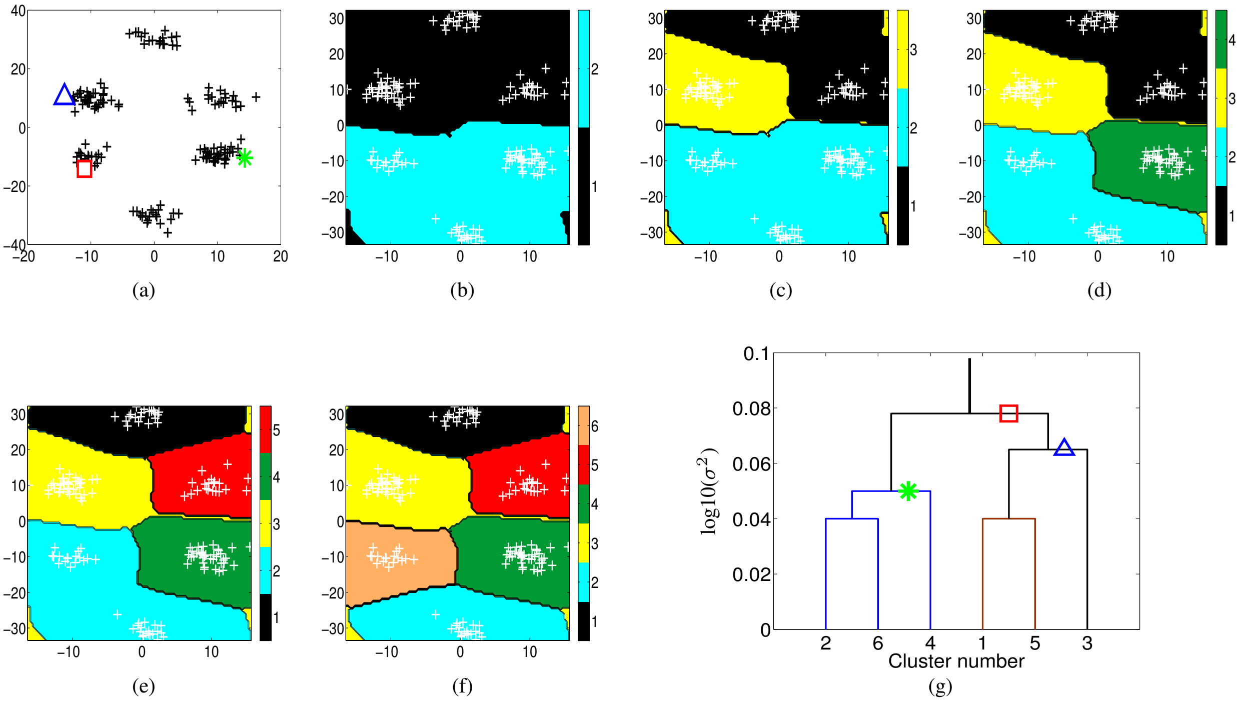
Image segmentation is a difficult task for spectral clustering due to the fact that the number of data points is large and therefore leading to eigen-decomposition of a big matrix. As it has been shown in [16], by incorporating side-information (labels in this case) to the unsupervised model, it is possible to improve the result of the unsupervised algorithm. Here we show how one can incorporate the labels in different levels of hierarchy from top to bottom to discover the multiscale structure of the given image. Experimental results on some color images from the Berkeley image data set [21] are shown in Fig. 5. For each image, a local color histogram with a 5×5 local window around each pixel is computed using minimum variance color quantization of eight levels. A subset of 500 unlabeled pixels together with some labeled pixels (see Table I) are used for training and the whole image for testing.

TABLE I

NUMBER OF LABELED AND UNLABELED DATA POINTS USED FOR TRAINING AND VALIDATION OF THE MODEL. Q IS THE NUMBER OF CLASS LABELS.

Image ID	Q	$\mathcal{D}$		$\mathcal{D}^{val}$	
		$\mathcal{D}_u$	$\mathcal{D}_L$	$\mathcal{D}_u$	$\mathcal{D}_L$
295087	3	500	8	3000	5
25098	3	500	6	3000	3
153072	3	500	14	3000	4
151087	3	500	6	3000	4

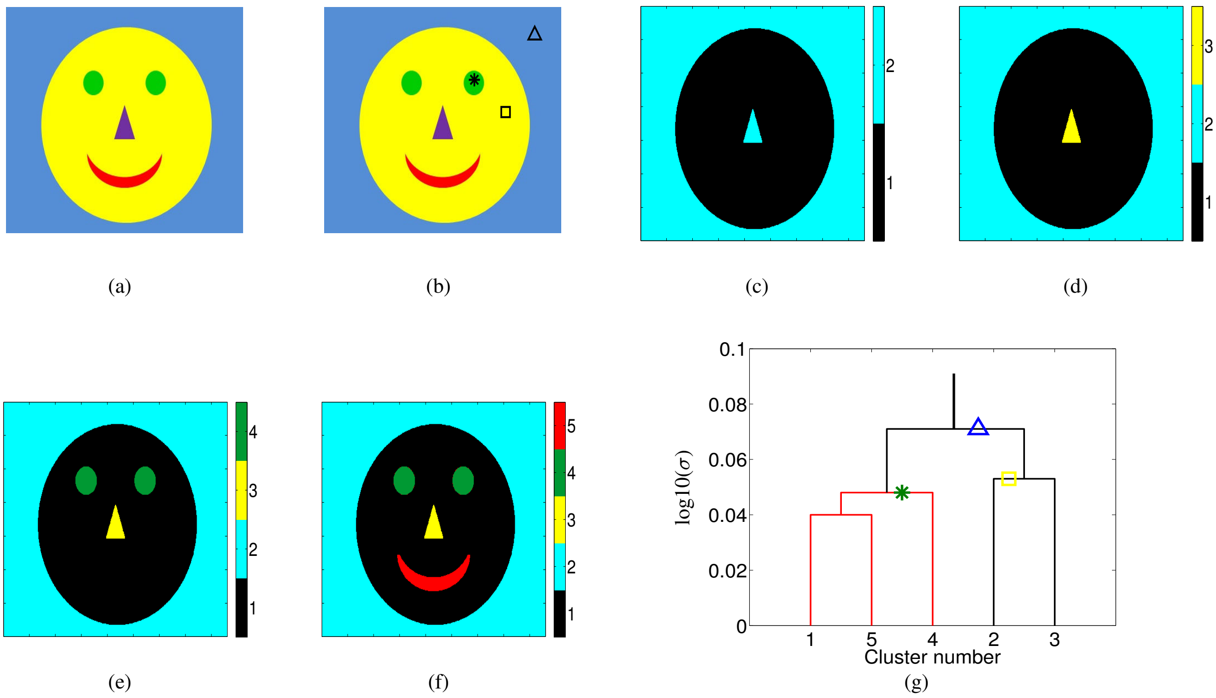
A comparison of the proposed method against classical hierarchical clustering, Linkage methods, including Average and Median Linkage (AL and ML), in terms of F-score is shown in Table II. Here the F-score is defined as F-score $= \frac{ARI}{DB}$ where ARI stands for the Adjusted Rand Index [22] and DB is the Davies Bouldin index [23]. Higher F-score values correspond to better quality clusters. It should be noted that the proposed HMSS-KSC is trained on a subset of the data set and the memberships of the entire data set is predicated based on the trained model. Whereas average and median linkage use the whole data set for obtaining the hierarchical results. The ground truth segmentations



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

(e) (f) (g)

Fig. 2. (a): Labeled data set used for the HMSS-KSC algorithm, labels are denoted by $*, \triangle, \square$ (b-f): Segmented data sets in different levels of hierarchy using the proposed approach. (g): The evolution of the clusters for top-down hierarchy.

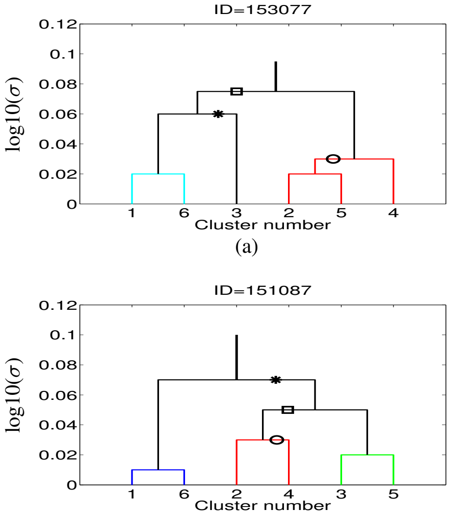


(a) (b) (c) (d)

(e) (f) (g)

Fig. 3. (a): Original image (b): Labeled image used in the HMSS-KSC algorithm, labels are denoted by $*, \triangle, \square$ (c-f): Segmented image in different levels of hierarchy using the proposed approach. (g): The evolution of the clusters for top-down hierarchy.

(human segmentation) of the images with different levels are provided in [21]. We have compared the obtained levels of hierarchy of the three algorithms with all of the available ground truth levels. The maximum F-score for each of the hierarchical levels is reported in Table II which shows the superiority of the HMSS-KSC compared to AL and ML algorithms even though a subset of a given image is used for training the HMSS-KSC model. In addition HMSS-KSC capitalizes on the labels incorporated at different levels of hierarchy.



(a)

(b)

Fig. 4. The evolution of the clusters for top-down hierarchy for three images with ID=100007, 153077 and 151087.

VI. Conclusions

In this paper, a semi-supervised hierarchical clustering algorithm is proposed that is able to discover the relation between clusters at different levels of hierarchy by integrating the available side-information (labels) into the analysis. The method has the out-of-sample extension property making it applicable for large data sets, by training only on a subset of data sets and predict the memberships of the rest based on the trained model. The algorithm is tested on real image data set and in most cases it outperforms the classical hierarchical clustering.

Acknowledgments

• EU: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC AdG A-DATADRIVE-B (290923). This paper reflects only the authors' views, the Union is not liable for any use that may be made of the contained information. • Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants • Flemish Government: ○ FWO: projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); PhD/Postdoc grants ○ IWT: projects: SBO POM (100031); PhD/Postdoc grants ○ iMinds Medical Information Technologies SBO 2014 • Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017). Johan Suykens is a professor at the KU Leuven, Belgium.

References

[1] X. Zhu, "Semi-supervised learning literature survey," *Computer Science, University of Wisconsin-Madison*, 2006.

[2] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 723–742, 2012.

[3] Y. Wang, S. Chen, and Z.-H. Zhou, "New semi-supervised classification method based on modified cluster assumption," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 5, pp. 689–702, 2012.

[4] S. Xiang, F. Nie, and C. Zhang, "Semi-supervised classification via local spline regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2039–2053, 2010.

[5] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *The Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[6] K. Bade and A. Nürnberger, "Creating a cluster hierarchy under constraints of a partially known hierarchy." in *SDM*. SIAM, 2008, pp. 13–24.

[7] K. Bade and A. Nurnberger, "Personalized hierarchical clustering," in *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*. IEEE, 2006, pp. 181–187.

[8] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[9] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[10] F. R. Chung, *Spectral graph theory*. AMS Bookstore, 1997, vol. 92.

[11] C. Alzate and J. A. K. Suykens, "Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 335–347, 2010.

[12] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least squares support vector machines*. Singapore: World Scientific Pub. Co., 2002.

[13] C. Alzate and J. A. K. Suykens, "Hierarchical kernel spectral clustering," *Neural Networks*, vol. 35, pp. 21–30, 2012.

[14] C. Alzate and J. A. K. Suykens, "A semi-supervised formulation to binary kernel spectral clustering," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2012, pp. 1992–1999.

[15] S. Mehrkanoon and J. A. K. Suykens, "Non-parallel semi-supervised classification based on kernel spectral clustering," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2012, pp. 2311–2318.

[16] S. Mehrkanoon, C. Alzate, M. Raghvendra, R. Langone, and J. A. K. Suykens, "Multi-class semi-supervised learning based upon kernel spectral clustering," *IEEE Transactions on Neural Networks and Learning Systems, In press*, 2014.

[17] S. Mehrkanoon and J. A. K. Suykens, "Large scale semi-supervised learning using ksc based model," in *Proceedings of the the 2014 IEEE World Congress on Computational Intelligence (IEEE WCCI/IJCNN 2014),July 6-11, 2014, Beijing, China*. IJCNN, 2014, pp. 4152–4159.

[18] V. Vapnik, *Statistical learning theory*. Wiley, 1998.

[19] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 28, no. 3, pp. 301–315, 1998.

[20] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, no. 1, pp. 53–65, 1987.

[21] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *in Proc. 8th International Conference on Computer Vision*, vol. 2. IEEE, 2001, pp. 416–423.

[22] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?" in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1073–1080.

[23] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus external cluster validation indexes," *International Journal of computers and communications*, vol. 5, no. 1, pp. 27–34, 2011.

TABLE II

COMPARISON OF HMSS-KSC, AVERAGE AND MEDIAN LINKAGE METHODS (AL, ML) FOR IMAGE SEGMENTATIONS IN TERMS OF F-SCORE INDEX

Ground Truth	295087			25098			153072			151087		
Level	HMSS-KSC	AL	ML	HMSS-KSC	AL	ML	HMSS-KSC	AL	ML	HMSS-KSC	AL	ML
1	**0.35**	0.15	0.13	0.17	**0.25**	0.18	**0.02**	0.01	0.01	0.16	0.14	**0.17**
2	**0.30**	0.16	0.21	0.10	0.04	**0.12**	**0.02**	0.01	0.01	**0.16**	0.14	0.13
3	**0.32**	0.14	0.14	0.08	0.08	0.08	**0.19**	0.16	0.17	**0.15**	0.13	0.14
4	**0.28**	0.12	0.10	0.08	0.08	0.06	**0.05**	0.03	0.03	**0.15**	0.12	0.13
5	**0.32**	0.13	0.14	0.07	0.07	0.07	**0.06**	0.05	0.04	**0.16**	0.14	0.12

Note: The higher the value of F-score the better the segmentation is.

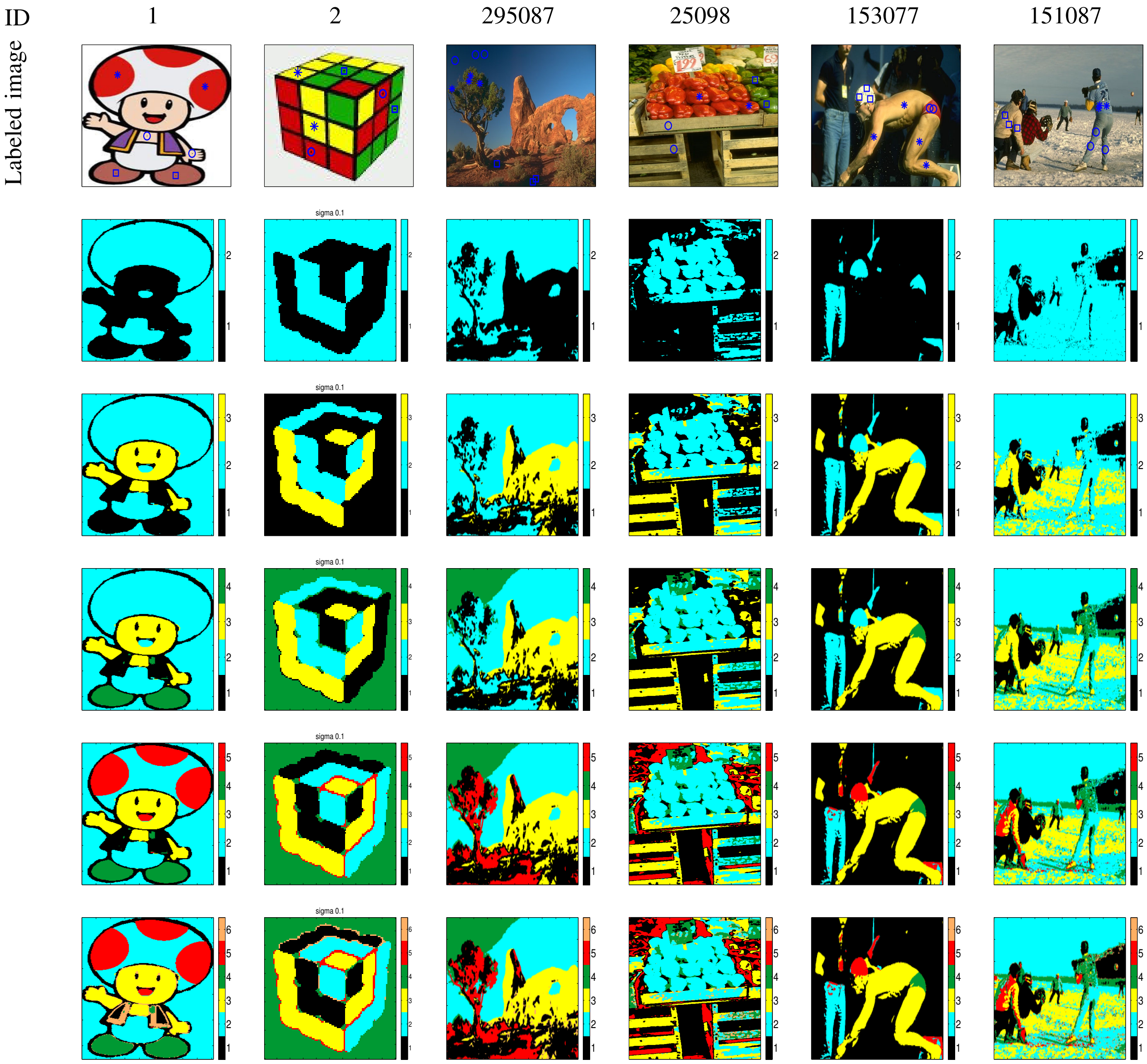


Fig. 5. Hierarchical image segmentation results using the proposed method. A subset of 500 and 3000 randomly chosen pixel histograms (unlabeled data points) together with some labeled data points are used for training and validation respectively. The whole image is used for testing. The labeled image is shown in the first row. The segmentation results obtained by the proposed HMSS-KSC for different levels of hierarchy are shown in the second-sixth rows.