STABLE CLUSTERING ENSEMBLE BASED ON EVIDENCE THEORY

Haijie Fu, Xiaodong Yue, Wei Liu, Thierry Denoeux

School of Computer Engineering and Science, Shanghai University, Shanghai, China
fhttsfhj@shu.edu.cn, yswantfly@shu.edu.cn
College of Electronic and Information Engineering, Tongji University, Shanghai, China
ldachuan@outlook.com
Universitéde technologie de Compiègne, CNRS UMR 7253 Heudiasyc, Compiègne, France
Shanghai University, UTSEUS, Shanghai, China
thierry.denoeux@utc.fr

ABSTRACT

As an unsupervised ensemble learning strategy, clustering ensemble combines multiple base clusterings into a high-quality one and has achieved successful applications in image analysis and data mining. However, extant clustering ensemble methods are ineffective to handle the data uncertainty in clustering consensus process, which may mislead to poor clustering ensemble results. To tackle the problem, we propose a stable clustering ensemble (SCE) method based on evidence theory (Dempster–Shafer theory) in this paper. Specifically, we construct a belief function of cluster membership to measure the uncertainty and stability of data instances in clustering ensemble and thereby implement the stable clustering ensemble algorithm. We test the proposed stable clustering ensemble method in the tasks of structural data clustering and image segmentation. The experimental results validate the proposed method is effective to process the uncertain data and produce high-quality data clusterings.

Index Terms— Clustering ensemble, stability, evidence theory

1. INTRODUCTION

Clustering ensemble methods have been proposed to combine multiple base clusterings into a single one, called the consensus, which aims at producing a more accurate and robust clustering of data [1]. Clustering ensemble has been successfully applied in the areas of image analysis [2], computer vision [3], multimedia [4] and data mining [5].

Although previous research works have achieved a great progress, extant clustering ensemble methods ignored the uncertainty of data in the clustering consensus process, which may mislead to poor clustering ensemble results. Specifically, if a data instance is partitioned unsteadily into different clusters in multiple base clusterings, it is indicated that the data instance is unstable in the consensus process and has much uncertainty in clustering ensemble. For the tasks of image analysis, the unstable pixels in the clustering ensemble correspond the uncertain regions of images. To improve the performances of data analysis, it is required to formulate and handle the data uncertainty in clustering ensemble.

To tackle the problem, we propose a stable clustering ensemble (SCE) method based on evidence theory (Dempster–Shafer theory) in this paper. In the proposed method, we construct a belief function of cluster membership to measure the uncertainty and stability of data instances in clustering ensemble and thereby implement the stable clustering ensemble algorithm. Based on the stability measure, we can divide data instances into two categories: *cluster core* and *cluster halo* [6]. Cluster core consists of the stable data instances with little uncertainty, which represent the structure of data distribution. In contrast, cluster halo contains the unstable data instances with high uncertainty, which denote the uncertain boundaries of clusters. To implement the stable clustering ensemble, we first detect the cluster core in clustering consensus as the certain fundamental clusters and then gradually distribute the uncertain data instances in the cluster halo into certain clusters. The main contributions of this paper are summarized below.

- *Propose a measure of data stability in clustering ensemble based on evidence theory.*
- *Implement a stable clustering ensemble algorithm with stability measure to handle uncertain data in clustering consensus.*

The rest of the paper is organized as follows. Section 2 briefly reviews the foundations of clustering ensemble and evidence theory. In Section 3, we introduce the proposed stable clustering ensemble method in detail, which include the stability measure with evidence theory and the clustering ensemble algorithm. In Section 4, the experiments are conducted to verify the proposed method. Section 5 concludes the paper work.

This work was supported by National Natural Science Foundation of China (Serial Nos. 61976134, 61991410, 61991415) and Natural Science Foundation of Shanghai (Serial No. 21ZR1423900).

2. BACKGROUND

Basics of Clustering Ensemble

Let $X = \{x_1, ..., x_n\}$ be a data set and $\Pi = \{\pi_1, ..., \pi_B\}$ denotes B base clusterings generated by multiple clustering procedures. Clustering ensemble aims at combining multiple base clusterings Π into an accurate and robust clustering. In general, clustering ensemble consists of two stages of generating base clusterings and assembling the consensus clustering result.

The diversity and quality of base clusterings are the key factors to affect the performances of clustering ensemble. Three kinds of strategies were investigated to generate diverse and accurate base clusterings, which include: 1) diverse parameter setting strategy [7] that uses one clustering algorithm with random cluster centers or cluster numbers; 2) diverse algorithm strategy [8] that uses various clustering algorithms to generate diverse base clusterings; 3) diverse feature strategy [9] that represents data clusterings in different feature spaces. For assembling the base clusterings, the methodologies can be categorized into following four kinds [10]. Feature based approach that transforms the clustering ensemble problem into the clustering of categorical data [11]. Direct approach, which is based on the relabeling process to find the best matched clustering [12]. Graph-based approach that utilizes the graph representation to solve the clustering ensemble problem [13]. Co-association approach that creates the pairwise correlation matrix among data instances to assemble base clusterings [14].

Preliminaries of Evidence Theory

Evidence theory, also referred to as Dempster–Shafer (D-S) theory or theory of belief functions [15, 16] is a theoretical framework for reasoning with partial and unreliable information. Let a variable w taking values in a finite set Ω, a mass function on Ω is defined as a mapping from 2^{Ω} to [0,1], satisfying the following condition

$$\sum_{A \in \Omega} m(A) = 1. \tag{1}$$

Each quantity $m(A)$ can be interpreted as the probability that the evidence supports $w \in A$. In particular, $m(\Omega)$ is the probability that the evidence tells us nothing about w, i.e., the unknown probability. A subset A of Ω such that $m(A) > 0$ is called a focal set of m. The mass function for which Ω is the only focal set is said to be vacuous, it represents total ignorance. Given a mass function m, belief bel and plausibility function pl are defined by

$$bel(A) = \sum_{\phi \neq B \subseteq A} m(B), \tag{2}$$

$$pl(A) = \sum_{B \cap A \neq \phi} m(B). \tag{3}$$

For all $A \subseteq \Omega$, the quantities $bel(A)$ and $pl(A)$ denote the degree of total support in A and the degree that the evidence consistent with A, respectively.

3. METHODS

3.1. Measuring data stability in clustering ensemble

Given B base clusterings for ensemble $\Pi = \{\pi_1, ..., \pi_B\}$, the uncertainty of a data instance in the clustering ensemble is related to the stability of the data instance belonging to a cluster under the partitions of different base clusterings [17]. Suppose some data instances are partitioned into a cluster in one base clustering, but in other base clusterings, these data instances are distributed into different clusters. It is natural to consider that the hypothesis of these data instances belonging to the same cluster is inconclusive. Next we utilize the evidence theory to measure the stability of belongingness of data to clusters.

We define a discernment frame $\Omega = \{c, \neg c\}$ to discern whether data belong to a cluster c or not ($\neg c$). For a data instance x_i belonging to a cluster c^{π_g} in the base clustering π_g, its mass function of cluster membership can be defined as

$$m_i^{\pi_g}(A) = \begin{cases} \dfrac{\sum_{t \neq g}^{\pi_t \in \Pi} \sum_{x_j \in c^{\pi_g}} f(i,j,\pi_t)}{(B-1) \cdot |c^{\pi_g}|}, & A = \{c^{\pi_g}\} \\ 1 - \dfrac{\sum_{t \neq g}^{\pi_t \in \Pi} \sum_{x_j \in c^{\pi_g}} f(i,j,\pi_t)}{(B-1) \cdot |c^{\pi_g}|}, & A = \Omega \end{cases} \tag{4}$$

where $|c^{\pi_g}|$ is the number of data instances in the cluster c^{π_g}. $f(i,j,\pi_t)$ is defined by

$$f(i,j,\pi_t) = \begin{cases} 1, & x_i, x_j \; in \; the \; same \; cluster \; of \; \pi_t \\ 0, & otherwise. \end{cases} \tag{5}$$

Through accumulating the evidences of data co-occurrence in a cluster from other $B-1$ base clusterings, $m_i^{\pi_g}(\{c^{\pi_g}\})$ denotes the probability mass of the instance x_i certainly belonging to the cluster c^{π_g} in the base clustering π_g, and $m_i^{\pi_g}(\Omega)$ denotes the unknown mass (uncertainty).

Based on the mass function above, we can further formulate the pairwise relationship between data instances using Dempster's combination rule. Suppose $\Theta = \{s, \neg s\}$ is a frame of discernment, in which s denotes a pair of data instances belonging to the same cluster, and $\neg s$ means that they belong to different clusters. For a pair of data instances x_i and x_j, the belief and uncertainty about the hypothesis that x_i and x_j belong to the same cluster are defined below.

$$bel_{ij}^{\pi_g}(\{s\}) = m_i^{\pi_g}(\{c^{\pi_g}\}) \cdot m_j^{\pi_g}(\{c^{\pi_g}\}), \tag{6}$$

$$m_{ij}^{\pi_g}(\Theta) = m_i^{\pi_g}(\Omega) \cdot m_j^{\pi_g}(\Omega). \tag{7}$$

If x_i and x_j belong to different clusters, $m_{ij}^{\pi_g}(\{\neg s\}) = 1$, $bel_{ij}^{\pi_g}(\{s\})$ and $m_{ij}^{\pi_g}(\Theta)$ are both zero.

Considering all the base clustering, the average belief and uncertainty about that x_i and x_j belong to a same cluster are obtained by

$$bel_{ij}(\{s\}) = \frac{1}{B}\sum_{t=1}^{B} bel_{i,j}^{\pi_t}(\{s\}), \tag{8}$$

$$m_{ij}(\Theta) = \frac{1}{B}\sum_{t=1}^{B} m_{i,j}^{\pi_t}(\Theta). \tag{9}$$

In general, if a data instance is certainly assigned to a cluster by most base clusterings, we consider that the instance is stable in the clustering ensemble. Therefore, we measure the stability of a data instance x_i through accumulating its probability mass in all base clusterings.

$$Stability(x_i) = \sum_{\pi_t \in \Pi} m_i^{\pi_t}(\{c^{\pi_t}\}) \tag{10}$$

Moreover, based on the stability measure, we can find the highly stable data instances from data sets to form the cluster core. Here we adopt the average stability degree of all data instances as the threshold to select the stable data in clustering ensemble.

3.2. Clustering ensemble with stable data

Based on the stability measure of data instances, we can implement the stable clustering ensemble to handle the uncertainty in clustering consensus. The process of the stable clustering ensemble consists of the following two stages.

1. In the fist stage, we select the highly stable data instances as the *cluster core* and perform clustering of these stable data to capture the certain clusters of data distribution.

2. In the second stage, considering the remained unstable data as *cluster halo*, we distributed each data instance in the cluster halo into a cluster in the certain clustering obtained in the first stage.

For a selected stable instance x_i among n data instances, its belief and uncertainty of pairwise relationship belonging to the same cluster are defined in equation (8) and (9). We utilize the pairwise belief and uncertainty to form the feature vector $R(i)$ of x_i,

$$R(i) = \{bel_{i1}(\{s\}), ..., bel_{in}(\{s\}), m_{i1}(\Theta), ..., m_{in}(\Theta)\}. \tag{11}$$

For each pair of stable instances from cluster core x_i, x_j, we can measure their similarity using cosine measure and construct a evidential similarity matrix (ESM), in which each element is computed as

$$ESM_{ij} = \frac{< R(i), R(j) >}{\sqrt{< R(i), R(i) > \cdot < R(j), R(j) >}}. \tag{12}$$

Algorithm 1 Clustering ensemble with stable data

Input: Stability degrees $stability(x_i)$ of data instances $\{x_1..., x_i..., x_n\}$ among B base clusterings;
Output: Clusters of data instances;

```
Average the stability degree as the threshold T;
cluster core = {x_i | stability(x_i) > T, i = 1, ..., n};
cluster halo = {x_i | stability(x_i) ≤ T, i = 1, ..., n};
Construct the similarity matrix ESM of cluster core;
Use HC algorithm on ESM to form clusters C of stable data;
while |cluster halo| > 0 do
    Get a data instance from cluster halo, assign it into the nearest cluster in C;
    Update the cluster and remove the instance from the cluster halo;
end while
return the updated clusters.
```

Based on ESM of all the stable data instances, we simply utilize a hierarchical clustering (HC) algorithm [18] to form the clusters of cluster core. These clusters can be considered as the certain part of the final clustering result. For the remained unstable data instances in cluster halo, we assign each instance into its nearest cluster of stable data instances and update the cluster iteratively until all the unstable data instances are assigned. The process of the stable clustering ensemble is shown in Algorithm 1. Using the algorithm, we can obtain a clustering ensemble result $P = \{pc_1, ..., pc_{k'}\}$, in which pc_i denotes the ith cluster and k' is the cluster number. If it is required to further merge the clusters in P, we can adopt the pairwise similarity measure between clusters pc_i and pc_j and perform HC algorithm again to merge the clusters in P.

4. EXPERIMENTS

In the experiments, we implement two tests to validate the superiority of the proposed stable clustering ensemble (SCE) method. The first test aims to verify the effectiveness of the SCE method for data clustering, we perform the SCE method on 10 structural data sets including both synthetic data sets (Flame, 2d-3c-no123, Aggregation, Chainlink, Wingnut) and UCI data sets (Ecoli, Segmentation, Glass, Knowledge Modeling, Yeast), and compare the clustering results with other 6 representative clustering ensemble methods, which include WTQ [19], WCT [19], CSM [19], MCLA [1], CSPA [1] and HGBF [20]. In the second test, we validate the ability of SCE for image analysis, we utilize the SCE method for image segmentation on Berkeley Segmentation Dataset and compare the segmentation results produced by other kinds of clustering methods.

For the experiment implementation, K-means algorithm is used to generate base clusterings and the cluster number of each base clustering is set as $\sqrt{n}$, n is the data instance number. In the comparative experiments, we run each algorithm

10 times and present the average results. Besides, the decay factor parameter in the comparative methods WCT, WTQ are set to 0.9.

In the first test on structural data sets, we adopt the well-known criteria ARI [21] and NMI [1] to evaluate the clustering quality and generate 50 base clusterings for ensemble on each data set. Table 1 and 2 list the ARI and NMI evaluations of clustering results produced by all the comparative clustering ensemble methods. It is obvious that the proposed SCE method achieves the best performance.

Table 1. ARI evaluations of comparative clustering methods

Data sets	*MCLA*	*HBGF*	*CSPA*	*CSM*	*WTQ*	*WCT*	*SCE*
Flame	0.5190	0.4858	0.4514	0.7171	0.6511	0.8081	**0.8392**
2d-3c-no123	0.6045	0.5265	0.5547	0.8617	0.8132	0.9109	**0.9849**
Aggregation	0.5613	0.5173	0.528	0.9338	0.9563	0.971	**0.9920**
Chainlink	0.4276	0.2110	0.1962	0.1988	0.3100	0.3751	**0.4784**
Wingnut	0.8276	0.899	0.769	0.8807	0.8304	0.8501	**0.9843**
Glass	0.2250	0.1884	0.1523	0.2527	0.1968	0.2098	**0.2572**
Ecoli	0.3637	0.2941	0.2816	0.3625	0.4359	0.3619	**0.7540**
KM	0.1722	0.1863	0.1849	0.2654	0.2412	0.2574	**0.3006**
Yeast	0.0909	0.0695	0.0673	0.0997	0.1075	0.0919	**0.1490**
Segmentation	0.4210	0.4283	0.3958	0.4811	0.3577	0.3405	**0.4769**

Table 2. NMI evaluations of comparative clustering methods

Data sets	*MCLA*	*HBGF*	*CSPA*	*CSM*	*WTQ*	*WCT*	*SCE*
Flame	0.4784	0.4599	0.4295	0.6389	0.5638	0.7284	**0.7833**
2d-3c-no123	0.6337	0.5849	0.6253	0.8608	0.7944	0.8947	**0.9575**
Aggregation	0.7681	0.7246	0.7290	0.9666	0.9566	0.9803	**0.9884**
Chainlink	0.2891	0.3455	0.0842	0.4318	0.4769	**0.5939**	0.5025
Wingnut	0.7697	0.8463	0.6673	0.8267	0.7445	0.7656	**0.9478**
Glass	0.3498	0.3160	0.2836	0.3807	0.3570	0.3606	**0.3822**
Ecoli	0.5465	0.4938	0.4921	0.5822	0.5763	0.5680	**0.7130**
KM	0.2700	0.2956	0.2981	0.3704	0.3331	0.3528	**0.4142**
Yeast	0.2023	0.1794	0.1725	0.2092	0.2280	0.2121	**0.2507**
Segmentation	0.5397	0.5398	0.5095	0.6059	0.5882	0.5336	**0.6418**

Besides the structural data, we also test the SCE method on unstructural images. We perform the clustering ensemble methods on the images from Berkeley Segmentation Dataset for image segmentation. Comparing with traditional clustering ensemble methods, SCE can effectively fuse multiple clustering-based segmentation results and detect the uncertain regions based on the stability measure. To illustrate this, we generate 20 segmentations using Chan-Vese method [22] with random initial contours. The stability of each pixel in an image is calculated by equation (10), Fig.1 shows the uncertain image regions (marked by gray color) that consist of the unstable pixels in multiple segmentation results.

Table 3. Evaluations of image segmentation results

Methods	*PRI* ↑	*VOI* ↓	*GCE* ↓	*BDE* ↓
WTQ	71.10	3.42	43.44	15.49
WCT	71.38	3.40	42.88	15.36
CSM	71.34	3.40	42.83	15.57
MCLA	72.05	3.43	43.72	14.73
CSPA	71.10	3.54	45.35	15.47
HBGF	71.64	3.49	44.43	15.11
SCE	**75.81**	**3.07**	**36.95**	**13.81**

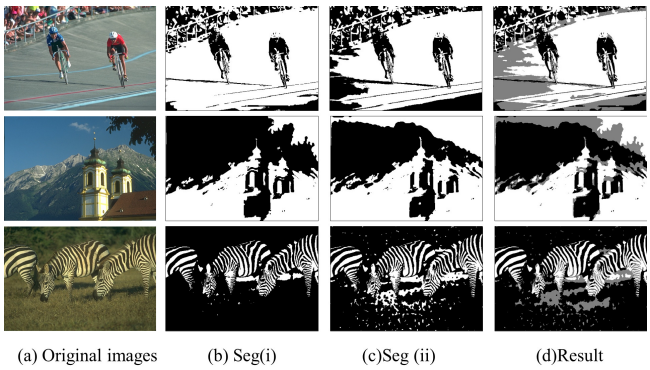

Fig. 1. Example of image segmentation based on SCE, (b-c) two different segmentation results, (d) the ensemble segmentation in which uncertain regions are marked by gray color.

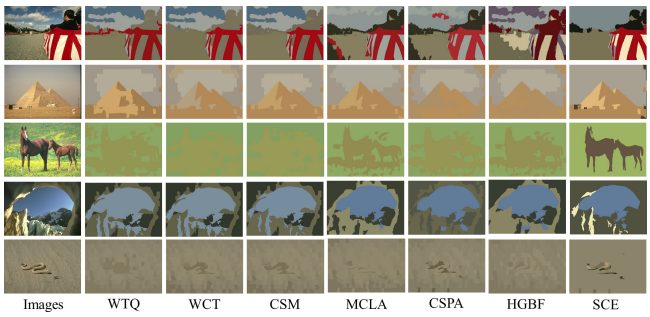

Fig. 2. Segmentation results based on different clustering ensemble methods.

Utilizing different clustering ensemble methods for image segmentation, we evaluate the clustering-based segmentation results by the measurements of BDE [23], PRI [24], VOI [25] and GCE [26]. To accelerate the clustering-based segmentation, we compress image pixels to superpixels [27] for clustering ensemble. Moreover, referring to [28], we initialize the cluster number from 2 to 6 for each image and determine the optimal cluster number to form the final image segmentation according to the highest PRI. Table 3 lists the detailed evaluations and Fig.2 presents some comparative segmentation results. We can find that our method produces more precise image segmentations than other clustering methods.

5. CONCLUSION

To tackle the drawback of handling the data uncertainty in clustering ensemble, we propose a stability measure of data in clustering ensemble based on evidence theory and thereby implement a stable clustering ensemble (SCE) method with the stability measure. The experiments of structural data clustering and image segmentation validate the superiority of the proposed SCE method. Our future work will focus on the acceleration of the stability computation.

6. REFERENCES

[1] Alexander Strehl and Joydeep Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of machine learning research*, vol. 3, no. Dec, pp. 583–617, 2002.

[2] X. Zhang, L. Jiao, F. Liu, L. Bo, and M. Gong, "Spectral clustering ensemble applied to sar image segmentation," *IEEE TGRS*, vol. 46, no. 7, pp. 2126–2136, 2008.

[3] M. Zhang, "Weighted clustering ensemble: A review," *Pattern Recognition*, p. 108428, 2021.

[4] X.D. Yue, D.Q. Miao, and L.B. Cao, "An efficient color quantization based on generic roughness measure," *Pattern Recognition*, vol. 47, no. 4, pp. 1777–1789, 2014.

[5] Alexander Topchy, A.K. Jain, and William Punch, "Clustering ensembles: Models of consensus and weak partitions," *IEEE TPAMI*, vol. 27, no. 12, pp. 1866–1881, 2005.

[6] Alex Rodriguez and Alessandro Laio, "Clustering by fast search and find of density peaks," *science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[7] H.G. Ayad and M.S. Kamel, "Cumulative voting consensus method for partitions with variable number of clusters," *IEEE TPAMI*, vol. 30, no. 1, pp. 160–173, 2007.

[8] Y. Yang and K. Chen, "Temporal data clustering via weighted clustering ensemble with different representations," *IEEE TKDE*, vol. 23, no. 2, pp. 307–320, 2010.

[9] Z.Q. Tao, H.F. Liu, S. Li, Z.M. Ding, and Y. Fu, "From ensemble clustering to multi-view clustering," in *IJCAI*, 2017.

[10] Natthakan Iam-On and Tossapon Boongoen, "Comparative study of matrix refinement approaches for ensemble clustering," *Machine Learning*, vol. 98, no. 1, pp. 269–300, 2015.

[11] Claudio Carpineto and Giovanni Romano, "Consensus clustering based on a new probabilistic rand index with application to subtopic retrieval," *IEEE TPAMI*, vol. 34, no. 12, pp. 2315–2326, 2012.

[12] B. Fischer and J.M. Buhmann, "Bagging for path-based clustering," *IEEE TPAMI*, vol. 25, no. 11, pp. 1411–1415, 2003.

[13] P. Zhou, X. Wang, and L. Du, "Clustering ensemble via structured hypergraph learning," *Information Fusion*, vol. 78, pp. 171–179, 2022.

[14] A. Fred and A.K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE TPAMI*, vol. 27, no. 6, pp. 835–850, 2005.

[15] G. Shafer, *A mathematical theory of evidence*, Princeton university press, 1976.

[16] T. Denoeux, Z. Younes, and F. Abdallah, "Representing uncertainty on set-valued variables using belief functions," *Artificial Intelligence*, vol. 174, no. 7, pp. 479–499, 2010.

[17] F. Li, Y. Qian, J. Wang, C. Dang, and J. Liang, "Clustering ensemble based on sample stability," *Artificial Intelligence*, vol. 273, pp. 37–55, 2019.

[18] S.C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[19] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, "A link-based approach to the cluster ensemble problem," *IEEE TPAMI*, vol. 33, no. 12, pp. 2396–2409, 2011.

[20] X.Z. Fern and C.E. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning," in *ICML*, 2004, p. 36.

[21] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[22] T.F. Chan and L.A. Vese, "Active contours without edges," *IEEE TIP*, vol. 10, no. 2, pp. 266–277, 2001.

[23] Jordi Freixenet, Xavier Munoz, David Raba, Joan Martí, and Xavier Cufí, "Yet another survey on image segmentation: Region and boundary information integration," in *ECCV*, 2002, pp. 408–422.

[24] Ranjith Unnikrishnan, Caroline Pantofaru, and Martial Hebert, "Toward objective evaluation of image segmentation algorithms," *IEEE TPAMI*, vol. 29, no. 6, pp. 929–944, 2007.

[25] Marina Meilă, "Comparing clusterings: an axiomatic view," in *ICML*, 2005, pp. 577–584.

[26] 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 *ICCV*, 2001, vol. 2, pp. 416–423.

[27] Radhakrishna Achanta and Sabine Susstrunk, "Superpixels and polygons using simple non-iterative clustering," in *CVPR*, 2017, pp. 4651–4660.

[28] T. Lei, X. Jia, Y. Zhang, S. Liu, H. Meng, and A.K. Nandi, "Superpixel-based fast fuzzy c-means clustering for color image segmentation," *IEEE Transactions on Fuzzy Systems*, pp. 1–1, 2018.