Proceedings of 2024 IEEE 12th International Conference on Intelligent Systems (IS)

Hierarchical Vs Centroid-Based Constraint Clustering for Animal Video Data

Samuel L. Hennessey, Francis J. Williams
and Ludmila I. Kuncheva
School of Computer Science and Engineering
Bangor University
Bangor, UK
Email: {sml18vly, f.j.williams, l.i.kuncheva}@bangor.ac.uk

***Abstract*—The identification of animals from video footage is an important ecological pursuit. It presents a labour-intensive task, requiring experts to invest considerable time analysing video recordings. While identifying humans from video is a mainstream research quest, much less has been done for recognising animals' identities. This paper explores and contrasts the effectiveness of hierarchical and centroid-based constraint clustering methods across five manually annotated video datasets containing animals. We aim to determine the most suitable methodology for integrating into a fully autonomous pipeline for animal re-identification. Our experimental findings indicate that, contrary to expectation, online hierarchical constraint clustering surpasses centroid-based constrained clustering.**

***Keywords*—Constrained clustering; animal re-identification; hierarchical clustering; pairwise constraints; semi-supervised learning**

I. Introduction

The identification of animals from video footage is increasingly vital, particularly in the context of wildlife conservation and monitoring farmyard animals. Thus far, monitoring wildlife has been a labour-intensive task, necessitating experts to devote extensive hours to scrutinising video recordings. With the ongoing deployment of video cameras in environments such as farms and remote wilderness areas, alongside the continual progress in species identification, the potential for establishing a fully autonomous animal re-identification pipeline is increasingly within reach.

In such a pipeline, the video input needs to be split into windows, and a suitable online clustering algorithm needs to be applied in order to create and maintain a description of the animals' identities.

The inclusion of constraints into clustering algorithms (Constrained Clustering (CC)) has proven to provide more accurate and robust results [1], [2], [3], especially in cases where the inherent structure of the data is complex or ambiguous. In conventional experimental setups, constraints are often derived from the ground truth labels of the data. In real-world situations, constraints may stem from expert opinions [2], [4]. However, constraints in video data are inherently generated from the videos themselves. Pair-wise constraints dictate whether two points should be grouped together (Must-Link) or kept separate (Cannot-Link).

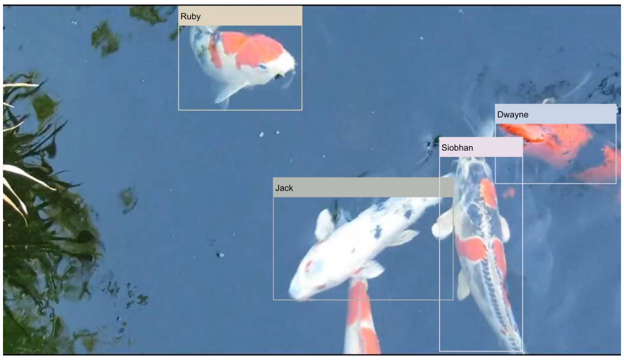
Centroid-based constrained clustering methods have proven their worth in experimental research [5], [6], [7], [8]. On the other hand, hierarchical constrained clustering methods appeared to have been less successful. However, video data present a challenge to centroid-based methods, rooted in the following argument. Clusters belonging to the same identity (same animal) may lie in far-apart regions of the feature space. Points will tend to be tightly packed together when they represent the same animal in contingent frames in the video. The animal's appearance will not change much from one frame to the next, leading to a string of very close points in the feature space. On the other hand, suppose that the animal disappears from camera view, and re-enters at a later stage, from a different view angle. The re-appearance will create another sequence of close points, possibly in a different part of the feature space. Figure 1 demonstrates our point.

We plotted one of the datasets used in the experiments in this study – the Koi fish video – in the two-dimensional space spanned by the first two t-sne[1] components of the data representation. It can be seen that points representing the same fish are distributed as pockets of non-spherically shaped clusters in different parts of the feature space. This suggests that centroid-based clustering methods may be inferior to hierarchical methods for the task of video data clustering.

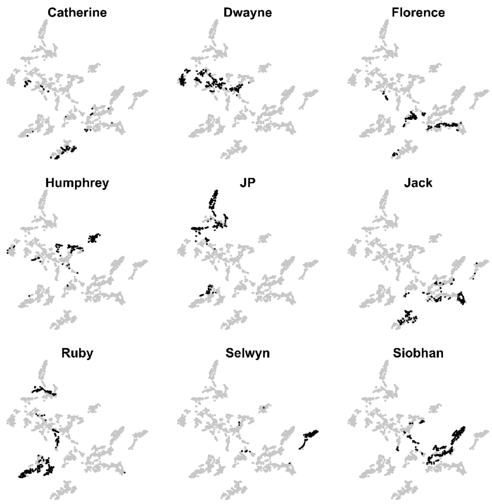
The novelty of this study lies in the comparison of centroid-based constrained clustering (CC) and hierarchical CC on an animal video data collection, to ascertain that hierarchical CC should be taken forward as the preferred candidate for a future fully autonomous animal re-identification pipeline. To this end, we experimented with different window sizes, using computationally inexpensive hierarchical CC methods.

The rest of the paper is organised as follows. Section II discusses related CC methodologies. Section III details the methods used in the comparative study. Section IV describes our experimental setup and Section V shows the results. Finally, we offer our conclusions in Section VI.

[1]t-sne (t-Distributed Stochastic Neighbor Embedding) is a dimensionality reduction method particularly useful for the visualising high-dimensional data. [9]



(a) An example of a frame from the annotated Koi fish video.



(b) Scatterplot of the data of the Koi fish video, in the first two t-sne components. Each true class label (fish identity) has a bespoke subplot, where the points for this class are shown in black.

Fig. 1: An example of distributed non-spherical clusters that are likely to occur in animal identification from video.

II. Related Work

In the present day where climate change and the environment are pressing issues, efficient animal monitoring methods have become more sought after [10]. This poses various challenges, especially when there is a large area or number of animals to cover. Current monitoring practises mainly rely on tagging or marking the animal in some form [11]. This is intrusive and may skew the findings as this distresses the animal. Non-intrusive methods, such as camera traps, have been used to instead capture video data from a secluded distance [12]. This solves half the problem, as the data still have to be annotated by hand. This manual analysis of video data is naturally time-consuming, cumbersome and prone to human error.

Object reidentification (re-id) in videos, e.g. Multiple Object Tracking (MOT), is a solution to this problem. MOT helps automate the data annotation process. Unfortunately, the majority of methods cover person and vehicle re-id, as well as using deep learning approaches [13], [14], [15]. The latter makes our task of animal re-id more difficult. This is because training a new model would require large amounts of annotated data, which is not easily obtainable or limited to only one species [16].

Clustering is an alternative that foregoes the need for such a model. One such method is hierarchical clustering, which combines points according to a distance metric until the desired number of partitions is achieved [17]. The suitability of the hierarchical approach to video data clustering is that the data is continuous. This means that between observations, i.e. between frames, we can expect the data to change minimally - resulting in string-shaped clusters.

The clustering process can be aided by utilising information in addition to the data representation, in the form of constraints [18]. Pairwise Constraints such as must-link (ML) and cannot-link (CL) offer insight into the clustering algorithm as to whether two points belong to the same cluster or not. Such incorporation of background information has shown to increase the accuracy of cluster methods [19].

Online Constrained Clustering (OCC) has been developed as a means to classify stream data. This is where data are presented to the clustering algorithm along a time line, rather than as a single batch. This creates a challenge of how to incorporate new information without having to re-cluster all of the data. One approach of OCC is through the use of windows. Here, the dataset is split into groups of a predetermined number of adjacent timestamps, such as a shot from a video [20]. The window application is better than 'plain' CC as only a small difference in the feature space is expected in adjacent observations. This means that objects in the window are expected to be closer in proximity to others in their class, as opposed to the rest of the dataset, and thus more likely to be grouped correctly. It should be noted that this is not full OCC as that involves propagating the class labels across the windows. For the purpose of this study we only consider small windows of adjacent frames.

III. Methods

1) Constrained Agglomerative Hierarchical clustering: Klein et al. [21] proposed that the integration of constraints into hierarchical clustering methods does not necessarily require modifications to the unsupervised algorithm itself. The majority of the hierarchical clustering methods can work from a distance matrix between all points in the dataset. Klein et al. [21] adjust the distance matrix to accommodate the constraints.

Initially, the must-link (ML) constraints may not represent a complete set of constraints. Pairs of objects missing from the original ML can be identified through transitive closure. To address this, ML is expanded to MLa, which includes all possible ML pairs. Initially, we construct a graph with N nodes, placing edges between all pairs of nodes corresponding to ML. Subsequently, we identify the connected components of the graph. For instance, if pairs (a, b) and (b, c) exist in ML, the connected component will include all three: a, b, and c. Finally, ML is augmented with all pairs in each connected component to arrive at MLa.

Next, consider a distance matrix $M_{N\times N} = a_{ij}$, where a_{ij} denotes the distance between points i and j. Constraints can be integrated into M by adjusting entry a_{ij} based on the constraint between points i and j. If $(i, j) \in MLa$, then $a_{ij} = 0$; conversely, if $(i, j) \in CL$, then $a_{ij} = \infty$.

Following the adaptation of the distance matrix to incorporate MLa and CL, unsupervised hierarchical clustering methods such as average linkage, single linkage, and complete linkage can be applied to a dataset to yield a partition. These respective variants will be referred to as `CAL` (Constrained Average Linkage), `CCL` (Constrained Complete Linkage), and `CSL` (Constrained Single Linkage) in the subsequent sections of this paper.

2) COP K-Means: Wagstaff et al. [2] introduced COP K-Means, a variant of the widely used centroid-based clustering method K-Means, designed to accommodate pairwise constraints such as Must-Link and Cannot-Link. Consider assigning a point x_i to a cluster. First, the clusters are ordered from the closest to the farthest. The next available closest cluster, C_j, is examined. The following two conditions are checked: (1) is there another point x_j such that $(x_i, x_j) \in ML$, and also $x_j \in C_k, C_k \neq C_j$ (violation of ML)? and (2), is there another point x_k such that $(x_i, x_k) \in CL$ and also $x_k \in C_j$? (violation of CL). If any of these conditions are met, x_i cannot be allocated to C_j, and the algorithm proceeds through the sorted list of clusters until finding one where x_i can be assigned without violating the constraints. If such a cluster does not exist, the algorithm halts. Since the order of points to be assigned to clusters is random, COP k-means often ends prematurely, without returning a viable solution.

3) Pairwise Confidence Constraints Clustering (PCCC): Beumann at al. [5] presented the PCCC algorithm, offering users the flexibility to specify pairwise constraints as either hard constraints (ML, CL), which must be strictly adhered to, or soft constraints (SML, SCL), where violations are permitted with associated penalties. The algorithm consists of five sequential steps: preprocessing, initialisation, assignment, update, and post-processing.

Prior to the algorithm's execution, the data is arranged as a weighted undirected graph $G = (V, E)$, where the vertices V correspond to the objects, and the edges E denote the constraints, categorised into four distinct groups: E^{ML} for hard must-link, E^{CL} for hard cannot-link, E^{SML} for soft must-link, and E^{SCL} for soft cannot-link. Notably, edges representing soft constraints are assigned weights denoted by a confidence value w_{ij}

In the preprocessing phase, the graph $G = (V, E)$ undergoes a transformation into another weighted undirected graph $G' = (V', E')$. This transformation involves contracting all edges $i, j \in E^{ML}$, merging nodes connected by hard must-link constraints, and adjusting edges to reflect hard cannot-link constraints, along with any remaining soft must-link and cannot-link constraints.

In the initialisation step, the initial positions of the k cluster centres are established, offering two distinct methods: either a random selection of points or the adoption of the K-Means++ algorithm introduced by Arthur et al. [22].

During the assignment step, every node in the graph $G' = (V', E')$ is allocated to one of the k clusters, aiming to minimise the total distance between nodes and their respective centres while adhering to both hard and soft pairwise constraints.

Following the assignment step, the positions of the cluster centres are adjusted based on the node assignments from the preceding step, a process iterated as long as there is potential for decreasing the objective function value. The assignment with the most favourable objective value upon termination is forwarded to the postprocessing step. Here, the labels of the graph $G' = (V', E')$ are remapped to the original representation $G = (V, E)$ and returned as the final assignment.

IV. Experimental Setup

A. Data

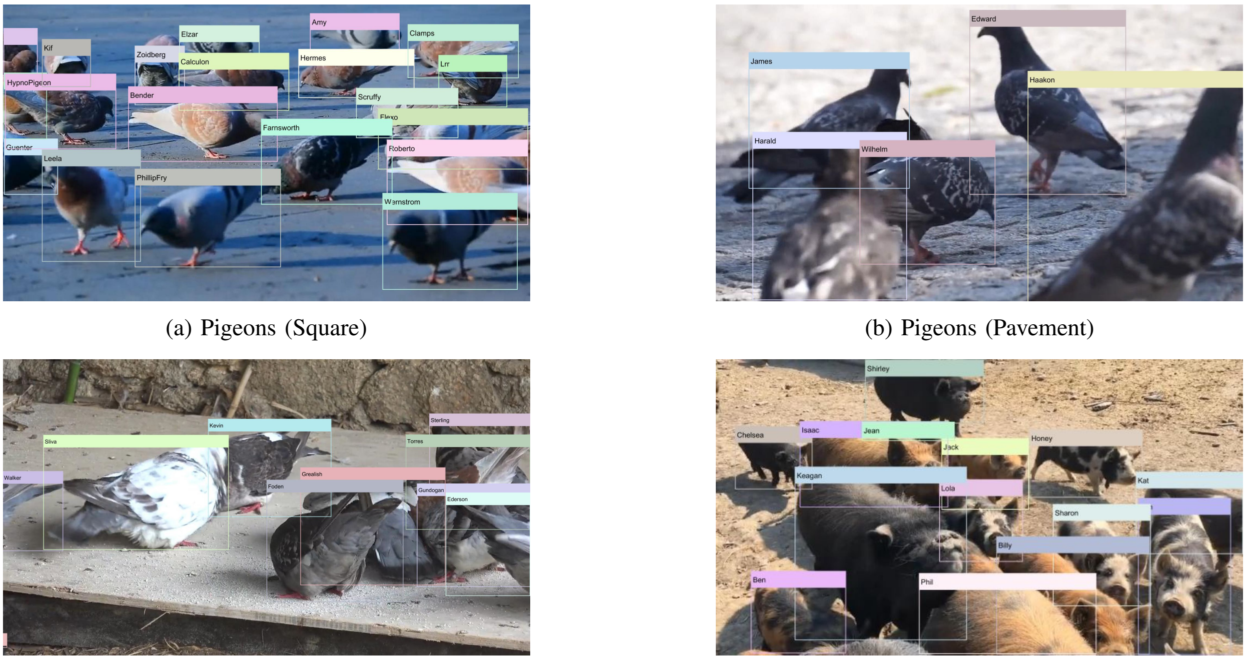
The dataset utilised for our case study comprises five video clips obtained from Pixabay (https://pixabay.com/) under the Pixabay Content License. The unconstrained videos capture the movement of groups of animals with durations ranging from 9 to 24 seconds. Each video features animals of the same species: Koi fish[2], pigeons (square)[3], pigeons (pavement)[4], pigeons (kerb)[5], and pigs[6], each of which are characterised in TableI. Each video has undergone manual annotation by creating bounding boxes (BB) containing one animal per box. These BBs have been assigned labels corresponding to the respective animal identities. Examples of annotated frames are depicted in Figure 1a for the Koi fish data, and in Figure 2 for the remaining four videos.

These videos are a part of our bespoke video dataset, which encompasses annotations, individual images, and files featuring various feature representations [23]. In this paper, we employed RGB moments (RGB) as the chosen feature representation as it was demonstrated in [24] to achieve the highest classification accuracy. To obtain the RGB feature representation of an image, you begin by dividing it into 3-by-3 blocks, subsequently, for each block we calculate the and store the mean and standard deviation of the red, green and blue panel, resulting in a total of 54 RGB features.

B. Metrics

Normalised Mutual Information (NMI) and Adjusted Rand Index (ARI) were employed to compare the obtained cluster

[2]www.pixabay.com/videos/koi-carp-fishes-ornamental-fish-5652/
[3]www.pixabay.com/videos/birds-street-pigeon-29033/
[4]www.pixabay.com/videos/pigeons-doves-and-pigeons-bird-city-4927/
[5]www.pixabay.com/videos/pigeons-eating-nature-birds-food-8234/
[6]www.pixabay.com/videos/pigs-farm-animals-livestock-49651/



(a) Pigeons (Square) (b) Pigeons (Pavement)

(c) Pigeons (Kerb) (d) Pigs

Fig. 2: Examples of annotated frames from the animal re-identification database used as our case study.

TABLE I: Characteristics of the videos

Video	k	l	N	c	Min p/f	Max p/f
Koi fish	536	22	1635	9	1	6
Pigeons (square)	300	9	4892	27	1	23
Pigeons (pavement)	600	24	3079	17	3	8
Pigeons (Kerb)	443	17	4700	14	8	13
Pigs	500	16	6184	26	4	20

Table notes: k is the number of frames; l is the video length in seconds; N is the number of objects (individual animal images); c is the number of classes (animal identities); Min p/f is the minimum number of animals per frame (image); Max p/f is the maximum number of animals per frame (image).

labels with the true labels of the data. An NMI score of 0 signifies no relationship between variables X and Y, while a score of 1 indicates a perfect relationship between the two variables. ARI yields a score between -1 and 1: a score close to 1 indicates a high degree of agreement between the clusterings, a score near 0 suggests random clustering, and negative scores imply disagreement. ARI is particularly suitable when the number of clusters varies or when the sizes of clusters differ.

C. Constraint Generation

Constraints for semi-supervised clustering algorithms are usually derived from samples of the ground truth labels of the dataset. However, when dealing with object data from videos, constraints can be naturally generated from the data itself. Cannot-Link constraints are created from each frame. No two objects in a frame can be in the same clusters, as they belong to different identities. For example, if we consider frame F_i and a list of objects $(O_1, ..., O_k)$ within F_i, we can create $k(k-1)/2$ cannot-link constraints.

Must-Link constraints are derived from pairs of consecutive frames (F_i, F_j). Each object is originally detected as a bounding box. We compute the overlapping region (Intersection over Union) between objects in successive frames. We expect objects from the first frame to be in a similar location in the subsequent frame. This allows us to determine, with high probability, if two objects are likely to be the same. We generate a must-link constraint between the object in F_i and the object in F_j if the intersection over union is above a threshold α. In this study we chose $\alpha = 0.7$.

D. Experimental Protocol

Each video is segmented into windows, where a window comprises a sequence of G consecutive frames. Our data sets contain the bounding boxes, their description as points in a multidimensional feature space, as well as the label (identity) for each bounding box. Constraints are generated for each window using the protocol outlined in Section IV-C. Once the constraints are established, each semi-supervised clustering method is applied to the window to create a partition. Then, we calculate the NMI and ARI between the resulting partition and the ground truth labels. Finally, we compute the average NMI and ARI values across all windows of size $G \in [2, 3, 4, 5, 10, 15, 20]$ for each clustering method and dataset.

Methods	Min WS (2)		Max WS (20)	
	NMI	ARI	NMI	ARI
klein02_SL	91.37%	91.23%	68.48%	69.64%
klein02_AL	91.41%	91.55%	70.68%	74.15%
klein02_CL	91.33%	91.24%	69.84%	72.82%
pccc	88.31%	86.63%	70.33%	73.35%
cop_kmeans	90.51%	89.77%	70.17%	73.93%

TABLE II: NMI and ARI values for the minimum window size of 2 and the maximum window size of 20, for all clustering methods used in the experiment averaged over all datasets.

V. Results

Table-II presents the normalized mutual information and adjusted Rand index for each clustering method, emphasising the minimum and maximum window sizes used in the experiment. The table clearly shows that using fewer frames per window leads to improved performance across all algorithms.

Figures 3, 4, 5, 6, and 7 illustrate the Adjusted Rand score for each clustering method applied to data segments (windows) with varying numbers of frames, with each plot corresponding to a different dataset. Figures 8a and 8b depict ARI and NMI of each clustering method averaged over all the datasets. It is evident from the plots that the number of frames within a window significantly influences the performance of the constraint clustering algorithms. By decreasing the number of frames within a window, we are able to reduce the complexity of the dataset, and provide the clustering algorithms an easier problem to solve. We observe that with a reduced number of frames, hierarchical methods outperform their centroid-based counterparts. This performance disparity is likely attributable to the inherent string-link structure produced by the moving of the objects within a window. However, as the number of frames per window increases, centroid-based algorithms begin to outperform hierarchical methods. Nevertheless, hierarchical methods with smaller window sizes still surpass the performance of centroid-based algorithms with larger window sizes.

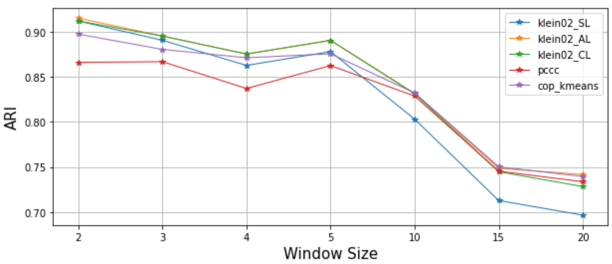


Fig. 3: ARI scores for each clustering method applied to segments of varying sizes on the koi dataset.

VI. Conclusion

In conclusion, the findings of this study underline the significance of optimising window size and leveraging appropriate clustering methodologies. We observe that the inherent structure and natural generation of constraints of video data

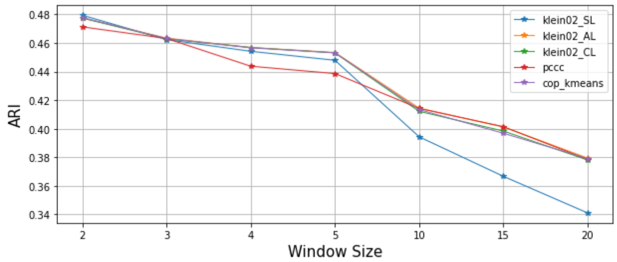


Fig. 4: ARI scores for each clustering method applied to segments of varying sizes on the pigeons (Kerb) dataset.

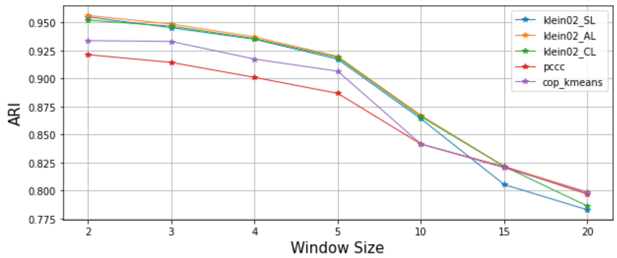


Fig. 5: ARI scores for each clustering method applied to segments of varying sizes on the pigeons (pavement) dataset.

are well-suited for hierarchical constraint clustering, and conducting clustering in a partially online manner significantly enhances the accuracy of the algorithms. Thus, performing online hierarchical constraint clustering within an animal re-identification pipeline is likely to yield the best results.

In future work, we plan to investigate label propagation between windows to achieve a global partition from online clustering. Our objective is to utilise the natural constraints to support and enhance this propagation.

References

[1] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *Proceedings of the 2004 SIAM international conference on data mining*. SIAM, 2004, pp. 333–344.

[2] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl *et al.*, "Constrained k-means clustering with background knowledge," in *Icml*, vol. 1, 2001, pp. 577–584.

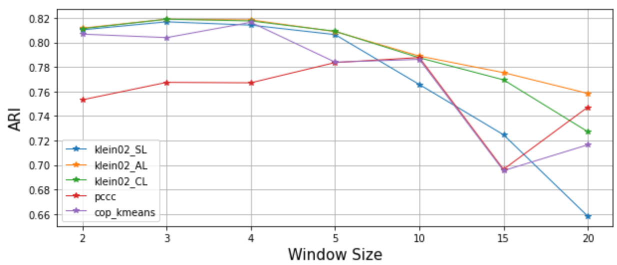


Fig. 6: ARI scores for each clustering method applied to segments of varying sizes on the pigeons (Square) dataset.

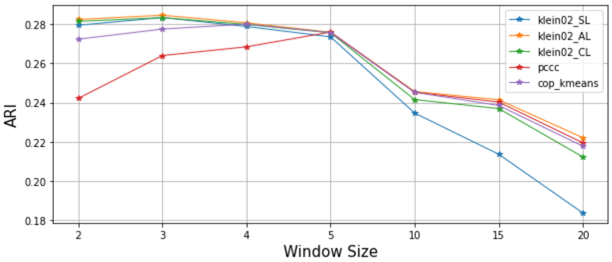
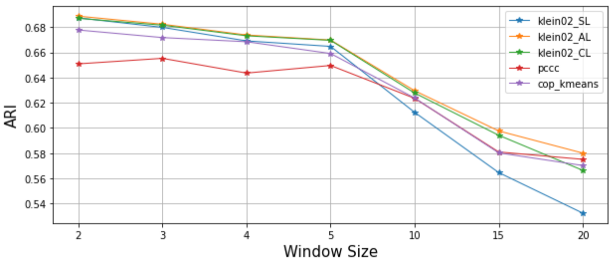
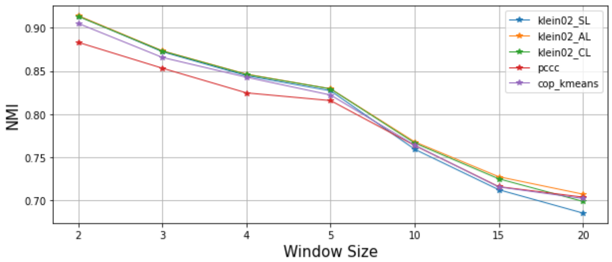


Fig. 7: ARI scores for each clustering method applied to segments of varying sizes on the pigs dataset.



(a) ARI - Windowed Data



(b) NMI - Windowed Data

Fig. 8: Metric scores for each clustering method applied to segments of varying sizes averaged over all the datasets.

[3] I. Davidson and S. Ravi, "Agglomerative hierarchical clustering with constraints: Theoretical and empirical results," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2005, pp. 59–70.

[4] S. Basu, M. Bilenko, and R. J. Mooney, "A probabilistic framework for semi-supervised clustering," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 59–68.

[5] P. Baumann and D. S. Hochbaum, "Pccc: The pairwise-confidence-constraints-clustering algorithm," *arXiv preprint arXiv:2212.14437*, 2022.

[6] L. I. Kuncheva, J. L. Garrido-Labrador, I. Ramos-Perez, S. L. Hennessey, and J. J. Rodríguez, "Semi-supervised classification with pairwise constraints: A case study on animal identification from video," *Information Fusion*, vol. 104, p. 102188, 2024.

[7] S. Overington, D. Edwards, P. Trinkl, and A. Buckley, "Application of constrained k-means algorithm for phase identification," in *2021 31st Australasian Universities Power Engineering Conference (AUPEC)*. IEEE, 2021, pp. 1–6.

[8] G. Hu, S. Zhou, J. Guan, and X. Hu, "Towards effective document clustering: A constrained k-means based approach," *Information Processing & Management*, vol. 44, no. 4, pp. 1397–1409, 2008.

[9] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[10] P. C. Ravoor and T. Sudarshan, "Deep learning methods for multi-species animal re-identification and tracking–a survey," *Computer Science Review*, vol. 38, p. 100289, 2020.

[11] W. E. Pine, K. H. Pollock, J. E. Hightower, T. J. Kwak, and J. A. Rice, "A review of tagging methods for estimating fish population size and components of mortality," *Fisheries*, vol. 28, no. 10, pp. 10–23, 2003.

[12] S. Schneider, G. W. Taylor, S. Linquist, and S. C. Kremer, "Past, present and future approaches using computer vision for animal re-identification from camera trap data," *Methods in Ecology and Evolution*, vol. 10, no. 4, pp. 461–470, 2019.

[13] G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61–88, 2020.

[14] L. Jiao, R. Zhang, F. Liu, S. Yang, B. Hou, L. Li, and X. Tang, "New generation deep learning for video object detection: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 8, pp. 3195–3215, 2021.

[15] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[16] M. Zuerl, R. Dirauf, F. Koeferl, N. Steinlein, J. Sueskind, D. Zanca, I. Brehm, L. v. Fersen, and B. Eskofier, "Polarbearvidid: A video-based re-identification benchmark dataset for polar bears," *Animals*, vol. 13, no. 5, p. 801, 2023.

[17] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[18] K. Wagstaff and C. Cardie, "Clustering with instance-level constraints," *AAAI/IAAI*, vol. 1097, pp. 577–584, 2000.

[19] G. González-Almagro, D. Peralta, E. De Poorter, J.-R. Cano, and S. García, "Semi-supervised constrained clustering: An in-depth overview, ranked taxonomy and future research directions," *arXiv*, 2023.

[20] P. Kulshreshtha and T. Guha, "An online algorithm for constrained face clustering in videos," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2670–2674.

[21] D. Klein, S. D. Kamvar, and C. D. Manning, "From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering," in *Proceedings of the Nineteenth International Conference on Machine Learning*, 2002, pp. 307–314.

[22] D. Arthur, S. Vassilvitskii *et al.*, "k-means++: The advantages of careful seeding," in *Soda*, vol. 7, 2007, pp. 1027–1035.

[23] L. I. Kuncheva, J. L. Garrido-Labrador, I. Ramos-Pérez, S. L. Hennessey, and J. J. Rodriguez, "Animal re-identification from video," Nov. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.7322821

[24] L. I. Kuncheva, J. L. Garrido-Labrador, I. Ramos-Pérez, S. L. Hennessey, and J. J. Rodríguez, "An experiment on animal re-identification from video," *Ecological Informatics*, vol. 74, p. 101994, 2023.