



UNIVERSIDAD NACIONAL DE COLOMBIA

A computational model for interpretable visual category discovery of foliar shapes

Modelo computacional para el descubrimiento interpretable de categorías de formas foliares

Jorge Enrique Victorino Guzmán

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2023

A computational model for interpretable visual category discovery of foliar shapes

Jorge Enrique Victorino Guzmán

Thesis submitted as requirement to obtain a:
Ph.D. en Ingeniería de Sistemas y Computación

Advisor:
Francisco Albeiro Gómez Jaramillo, Ph.D.

Computational Modelling of Biological Systems (COMBIOS)

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2023

Dedication

To my family

To my mother, who has always been my rock and my inspiration. To my wife, who has supported me through every step of this journey and always believed in me. And to my son Sebastian, the light of my life and the reason I strive to be better every day. This thesis is dedicated to all of you, with love and gratitude.

Acknowledgements

I want to express my deepest gratitude to my thesis advisor, Francisco, for his dedication and invaluable knowledge that he has shared with me throughout this process. His guidance and support have been instrumental in the completion of this thesis.

I also want to extend my thanks to the Universidad Central, and its Programa de Desarrollo Profesorado, for providing me with the resources and support needed to complete this research. I am grateful for the opportunity to work with such a renowned institution and for the support provided by the faculty and staff throughout my studies.

I am forever grateful for all the support and hard work of my advisor, Francisco, and the University, for helping me make this thesis a reality.

Publications

Publications

Journals

- **Victorino, J.**, and Gómez, F. (2019). “Contour analysis for interpretable leaf shape category discovery”. *Plant methods*, **15**(1), 1-12.
- **Victorino, J.**, Rudas, J., Reyes, A.M., Pulido, C., Chaparro, L.F., Estrada, C., Narvaez, L.A. and Gómez, F. (2021). “Highly Sessional Aggressive Behaviors Link to Temporal Dynamics Shared Across Space”. *IEEE Access*, **9**, 165072-165084.
- Chaparro, L. F., Pulido, C., Rudas, J., **Victorino, J.**, Reyes, A.M., Estrada, C., Narvaez, L.A. and Gómez, F. (2021). “Quantifying Perception of Security Through Social Media and Its Relationship With Crime”. *IEEE Access*, **9**, 139201-139213.
- Villamil, J., **Victorino, J.**, and Gómez, F. (2021). “The effect of mobile camera selection on the capacity to predict water turbidity”. *Water Science and Technology*, **84**(10-11), 2749-2759.

Proceedings

- **Victorino, J.**, Rudas, J., Reyes, A. M., Pulido, C., Chaparro, L. F., Narváez, L. Á., Martínez, D., and Gómez, F. (2020, December). “Spatial-temporal patterns of aggressive behaviors. A case study Bogotá, Colombia”. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 667-672). IEEE.
- **Victorino, J.**, Barrero, M., Rudas, J., Pulido, C., Chaparro, L., Estrada, C., Narváez, L. Á. and Gómez, F. (2022, February). “Prediction based on time series of aggressive behaviors. A case study Bogotá, Colombia”. In *2022 International Symposium on Electrical, Electronics and Information Engineering (ISEEIE)* (pp. 114-119). IEEE.
- Chaparro, L. F., Pulido, C., Rudas, J., Reyes, A. M., **Victorino, J.**, Narváez, L. Á., Martínez, D., and Gómez, F. (2020, December). “Sentiment analysis of social network content to characterize the perception of security”. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 685-691). IEEE.
- Pulido, C., Reyes, A. M., Rudas, J., **Victorino, J.**, Martínez, D., Narváez, L. Á., and Gómez, F. (2020, November). “An evolutionary algorithm for reducing fear of crime”.

In 2020 7th International Conference on Behavioural and Social Computing (BESC) (pp. 1-6). IEEE.

- Reyes, A. M., Rudas, J., Pulido, C., **Victorino, J.**, Martinez, D., Narváez, L. Á., and Gómez, F. (2020, November). “Characterization of temporal patterns in the occurrence of aggressive behaviors in Bogotá (Colombia)”. *In 2020 7th International Conference on behavioral and social computing (BESC)* (pp. 1-4). IEEE.
- Rudas, J., Reyes, A.M., Pulido, C., Chaparro, L.F., **Victorino, J.**, Narváez, L. Á., Martinez, D. and Gómez, F. (2020, December). “Consistent spatial decomposition of temporal occurrence of aggressive behaviors: A case study in Bogotá, Colombia”. *In 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 715-719). IEEE.
- Reyes, A. M., Rudas, J., Pulido, C., Chaparro, L. F., **Victorino, J.**, Narváez, L. Á., Martinez, D., and Gómez, F. (2021). “Multimodal prediction of aggressive behavior occurrence using a decision-level approach”. *In 11th International Conference of Pattern Recognition Systems (ICPRS 2021)* (pp 163 – 169).
- Chaparro, L.F., Pulido, C., Rudas, J., Reyes, A.M., **Victorino, J.**, Narváez, L. Á., Martinez, D., and Gómez, F., (2021, May). “Interpretability Of The Perception Of Security Based On Tweets Content”. *In 2021 International Conference on Applied Artificial Intelligence (ICAPAI)* (pp. 1-6). IEEE.
- Pulido, C., Chaparro, L. F., Rudas, J., **Victorino, J.**, Estrada, C., Narváez, L. Á., and Gómez, F. (2021, May). “Prediction of Perception of Security Using Social Media Content”. *In Iberoamerican Congress on Pattern Recognition* (pp. 88-96). Springer, Cham.

Conferences

- **Victorino J.** and Gómez, F., “A computational model for interpretable visual category discovery of foliar shapes”. *Sexto Coloquio doctoral Universidad nacional, Bogotá, Colombia, May 11 2018.*
- **Victorino, J.**, Rudas, J., Reyes, A. M., Pulido, C., Chaparro, L. F., Narváez, L. Á., Martinez, D., and Gómez, F. (2020, December). “Spatial-temporal patterns of aggressive behaviors. A case study Bogotá, Colombia”. *In 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, The Hague (Online), Netherlands, December 7 2020.
- **Victorino, J.**, Barrero, M., Rudas, J., Pulido, C., Chaparro, L., Estrada, C., Narváez, L. Á. and Gómez, F. “Prediction based on time series of aggressive behaviors. A case

study Bogotá, Colombia”. *In 2022 International Symposium on Electrical, Electronics and Information Engineering (ISEEIE)*, Hangzhou (Online), China, November 19 2021.

Abstract

Leaf morphological description studies are complex because they require highly trained personnel and the consultation of a large amount of available documentation, such as visual category systems in botanical manuals, books, online databases, and herbariums, and commonly should be contrasted with other experts. These studies require a significant resource investment and a high manual workload. On the other hand, the number of botanists available and in training for performing these studies cannot meet the current needs of the growing amount of foliar information resulting from automation and the increasing complexity of research questions. In this scenario, automatic computational processes are required to provide a qualitative and quantitative morphological description that significantly alleviates the experts' workload.

The difficulty of using automatic approaches in morphological analysis materializes if any of these functionalities are missing: 1. extracting the relevant features from the shape so that they can be analyzed separately, 2. producing robust categories that emerge from the representation of each feature, and 3. explanatory capacity of the categories in the context of the biological problem.

This work proposes a computational strategy for discovering leaf-shape categories that helps to overcome these limitations. First, an algorithm extracts each feature and represents it appropriately (contractive) in a specific feature space (morphospace). Then, the points in the morphospace are analyzed and organized under the concepts of neighborhood, cohesion, and persistence. The method accounts for these features and analyzes the number of clusters for all neighborhood sizes, and chooses the optimal number of clusters, in other words, the number of categories. This system of categories has the property of explaining the underlying phenomenon qualitatively and quantitatively. In this way, during the neighborhood analysis, the categorization dendrogram emerges. Finally, the interpretation of the results is given by the morphospace and by the dendrogram.

The effectiveness of the proposed approach is evaluated against category systems established by experts. The results show that the proposed approach can produce useful categorizations similar to what is reported in Hickey's manual, a widely used botanist manual. This approach allows biologists to make qualitative and quantitative descriptions of leaf morphology, helping them to describe variability, taxonomy, plasticity, adaptation, and ecological changes.

Keywords: (Novel category discovery, Unsupervised categorization, Leaf shape, Contour analysis, morphological, Image processing, Topological analysis, Image classification).

Resumen

Los estudios de descripción morfológica de hojas son complejos en la medida que requieren de personal altamente entrenado y de la consulta de una gran cantidad de documentación disponible como i.e., sistemas de categorías visuales en manuales botánicos, libros, bases de datos en línea, herbarios, inclusive contrastar hallazgos con otros expertos. Por tanto, estos estudios demandan una inversión significativa de recursos y tienen una alta carga de trabajo manual. Por otro lado, la cantidad de botánicos disponibles y en formación no logra suplir las necesidades actuales de la creciente cantidad de información foliar resultante de la automatización y la creciente complejidad de las preguntas de investigación. En este escenario se requieren procesos computacionales automáticos que proveen una descripción morfológica cualitativa y cuantitativa que alivian en gran medida la carga de trabajo de los expertos.

La dificultad de usar enfoques automáticos en análisis morfológicos se materializa si hace falta alguna de estas funcionalidades: 1. extraer los rasgos relevantes de la forma para que puedan analizarse por separado, 2. producir categorías robustas que emergen de la representación de cada rasgo, y 3. capacidad de explicación de las categorías en el contexto del problema biológico.

En este trabajo se propone una estrategia computacional para el descubrimiento de categorías de formas de hojas que ayuda a automatizar estas funcionalidades clave. Primero, un algoritmo extrae cada rasgo y lo representa de manera adecuada (contractiva) en un espacio de características (morfoespacio) específico. Luego, la muestra proyectada en el morfoespacio es analizada y organizada bajo los conceptos de vecindad, cohesión y persistencia. Este método realiza un análisis del número de grupos para todos los tamaños vecindad y escoge la cantidad de grupos óptima, en otras palabras, las categorías. Este sistema de categorías tiene la propiedad de explicar el fenómeno subyacente de manera cualitativa y cuantitativa. De esta forma, durante el análisis de vecindad surge el dendrograma de la categorización. La interpretación de los resultados está dada por el morfoespacio y por el dendrograma.

La efectividad del enfoque propuesto se evalúa frente a sistemas de categorías establecidos por expertos. Los resultados evidencian que el enfoque puede producir categorizaciones razonables similares a lo reportado en el manual de Hickey. Este enfoque permitirá a los biólogos hacer descripciones cualitativas y cuantitativas de la morfología útiles en estudios de variabilidad morfológica, taxonomía, plasticidad, adaptación y ecología.

Palabras clave: (Descubrimiento de categorías novedosas, categorización no supervisada, forma de hoja, análisis de contorno, morfología, procesamiento de imágenes, análisis topológico, clasificación de imágenes).

Esta tesis de doctorado se sustentó el 06 de 07 de 2023 a las 7:00 a.m. y fue evaluada por los siguientes jurados:

Lauren Raz (Phd.)
Profesor Asociado, Instituto de Ciencias Naturales
Universidad Nacional de Colombia

Mary Lee berdugo Lattke (Phd.)
Profesor Asistente, Facultad de Ingeniería
Universidad Central

Jorge Eliécer Camargo Mendoza (Phd.)
Profesor Asistente, Facultad de Ingeniería
Universidad Nacional de Colombia

Contents

Acknowledgements	vii
Abstract	xiii
Abbreviations and symbols	xxi
1 Introduction	1
2 Contour analysis for interpretable leaf shape category discovery	9
2.1 Background	9
2.2 Methods	11
2.2.1 Contour extraction	11
2.2.2 Contour representation	11
2.2.3 Dimensionality reduction	12
2.2.4 Clustering	13
2.2.5 Evaluation	14
2.3 Results	16
2.3.1 Capability of the method to recover the original categories	16
2.3.2 Qualitative evaluation	17
2.4 Discussion	20
2.5 Conclusions	25
3 Robust visual category system of specific leaf shape traits	26
3.1 Introduction	26
3.2 Related Work	29
3.3 Materials and methods	30
3.3.1 Data and reference visual categories	30
3.3.2 Leaf segmentation	33
3.3.3 Petiole segmentation	33
3.3.4 Contour extraction	33
3.3.5 Feature extraction	34
3.3.6 Category discovery algorithm	37
3.3.7 Experimental setting	44

3.4	Results	45
3.4.1	Discovering visual shape categories for leaf traits	45
3.4.2	Discovering species as categories	49
3.5	Discussion	51
3.5.1	Importance of finding specific trait categories	54
3.5.2	Importance of defining robust categories	55
3.5.3	Results interpretability	56
3.5.4	Limitations and future work	57
4	Conclusions and recomendations	58
4.1	Contributions	58
4.2	Conclusions	61
4.3	Limitations	62
	Bibliography	64

Abbreviations and symbols

Abbreviations

Abbreviate	Meaning
<i>AI</i>	Artificial Intelligence
<i>CFT</i>	Complex Fourier Transform
<i>DT</i>	Delaunay Triangulation
<i>MS</i>	Meanshift algorithm
<i>PC1</i>	First Principal Component
<i>PC2</i>	Second Principal Component
<i>PC3</i>	Third Principal Component
<i>PCA</i>	Principal Component Analysis
<i>SVD</i>	Singular Value Decomposition
<i>TDA</i>	Topological Data Analysis
<i>VD</i>	Voronoi Diagram
<i>W</i>	Whitening method

Symbols

Symbol	Term	Meaning
α_i	Filtration value	Value of scale distance
$D(P_i, P_j)$	Distance among two points	Value in the distance matrix (i, j)
$D_{i,j}$	Distance among two points	Value in the distance matrix (i, j)
$D(P)$	Delaunay triangulation in P	Delaunay triangulation of morphospace points

Symbol	Term	Meaning
δ_i	Density in point i	Density value in the point i position
f_i	Factor i	Value to compensate density for the point i
\mathcal{K}_i	Simplicial complex	Set of simplices for a given α_i
N	Number of samples	Number of samples in the experiment
q	Minimum samples per group	Minimum number of samples to define a group

1 Introduction

Since ancient times, studying plant morphology has been a matter of survival [58]. Identifying which plants are edible and which are toxic or poisonous for human consumption is essential [58]. Furthermore, the knowledge of identifying plants and associating them with their medicinal properties has always been highly esteemed and recognized in all cultures [102]. Hence, the need arose to recognize the different plants in the environment and study how they can be discriminated against, classified, and organized to refine and pass on this knowledge through generations.

A significant historical milestone in describing these biological objects is the contribution of Carl Linnaeus, who developed the classification system known as binomial nomenclature, revolutionizing the way species are categorized and named [39]. This categorization system enabled clear and precise communication in the field of biology and laid the foundation for modern taxonomy [102]. Since then, classification systems based on categories have been fundamental for describing biological objects.

For constructing these systems, botanical experts typically use various physical or morphological characteristics the plant provides, particularly the leaves, to identify and discriminate among them. These characteristics include size, texture, consistency, leaf venation pattern, and shape, among others [39]. Obtaining these features involves exploring many leaves, requiring physical access to the plant and a method to capture, collect, and store all this information. In addition, this data should be described for particular biological contexts, commonly by a morphological description of their features. The morphological description is a highly specialized pattern recognition task that commonly demands extensive knowledge of botany and advanced use of information from herbariums, online databases, and a widely used botanic manual, such as Hickey's manual [58]. Additionally, several years of field experience and interaction with other experts, often through research collaborations, are necessary to describe specific groups of plants [58].

In the case of plants, the knowledge resulting from this process is documented in botanical architecture manuals, where various relevant traits are categorized visually, including detailed descriptions and graphic definitions of botanical terms [128]. The observation that this process is heavily dependent on human expertise leads to the question of whether *it is computationally possible to reproduce botanists' specialized knowledge when identifying plant*

species or varieties and whether the mechanism of constructing visual categories used by biologists can be replicated algorithmically?. This thesis explores this research question in the context of the morphological description of leaves.

When defining shape categories, one issue is the inevitable introduction of biases [132]. These biases depend on the educational background of the expert and the shapes they have studied or encountered, making it impossible to have a complete system of categories covering all possible forms [132]. In most botanic manuals, some categories are defined qualitatively using simple shapes, while others are defined more quantitatively using measurements such as area, length, and angles [35]. These definitions may be ambiguous. For example, Hickey’s manual defines apex angle categories precisely. The apex is an “acute” class if the angle at the top is less than 90 degrees and “obtuse” in other cases. However, these defined thresholds may work well in a specific biological context but may not be suitable in others [60]. In the description of natural system properties, this ambiguity is commonly present. For instance, some values, like thermal floors, are also arbitrarily set by dividing altitudes into ranges such as “warm climate” between 0 and 1000 meters and “temperate climate” between 1000 and 2000 meters. These categorization systems can also lead to classification problems when elements fall close to the boundaries [110].

Moreover, defining the shape of a plant can be an ambiguous process as it depends on the subjective perception of the expert [90]. It has even been found that people can classify differently at different times. Objective systems should work consistently across different scenarios and situations. In addition, fixed reference systems need to be imported with the emergence or consideration of new forms [26].

In conclusion, plant morphological knowledge and identification have been fundamental for human survival and development throughout history [10]. Although experts have developed effective classification and categorization systems, these systems still have biases and subjectivity [132]. The possibility of computationally reproducing specialized knowledge and the mechanism of constructing visual categories is an ongoing challenge [3]. Nevertheless, technological advancements undoubtedly open new opportunities in this field.

Shape description and category discovery

Currently, biologists employ sophisticated methods for the qualitative description of shape and quantitative analysis through geometric morphometrics [73, 16]. For instance, Procrustes’s popular approach allows them to focus on specific shape traits, typically a segment of the overall contour, and test one or multiple morphological hypotheses [16]. Despite the popularity of this method in biology, this approach requires arduous and highly specialized

manual work, making it highly sensitive to placing critical points known as “landmarks” and “pseudo-landmarks”. These marks are crucial points that the expert manually places along the contour [20]. The precise positioning of landmarks is essential to obtain high-quality results that validate or reject the hypotheses [20]. This approach’s primary limitation lies in accurately reproducing all landmarks in each sample [3].

In Procrustes analysis, after registering all the landmarks in each sample, a dimensionality reduction using Principal component Analysis (PCA) is used, followed by a clustering technique such as k-means, where the number of groups, k , must be defined beforehand, to determine a set of categories [16]. This strategy is limited because, apriori, the number of groups has yet to be discovered. Therefore, more automated and efficient approaches are needed to handle large-scale data sets while maintaining accuracy and reliability [8]. These newer methods should alleviate the burden of manual landmark placement and reduce the risk of human error, providing more reproducible and consistent results [71].

An alternative approach involves automatically placing points uniformly along the entire contour, bypassing the need to set specific landmarks manually [3]. Subsequently, base functions allow to extract particular properties that facilitate data analysis [116]. This approach assesses how closely the shape resembles well-known functions, such as sines, complex cosines, wavelets, and Gaussians [21]. This approach results in representations known as Fourier ellipses, p-type, wavelets, Gabor filters, or chain codes, respectively [115].

In this last strategy, a morphospace is constructed by measuring the similarity of each sample to the selected base functions [128]. For example, using Fourier ellipses, the resemblance of a leaf’s representation to a cosine of a particular frequency is evaluated, exploring multiple frequencies to achieve a comprehensive frequency-domain analysis of the data [68]. Conversely, wavelet or Gabor functions are bounded, unlike sines and cosines, which are infinite. This property allows scaled and shifted versions of these functions to select or extract specific segments of the shape [128]. The goal here is to obtain representations of shape features that exhibit contractiveness. In this context, contractive means that samples projected in the morphospace become closer the more similar they are, and these similarity relationships remain consistent despite changes in scale, displacement, rotation, or other transformations [116].

Despite the flexibility of these characterizations, approaches using Fourier ellipses or p-type representations fail to distinguish between different shape features, leading to a loss of specificity [3]. This limitation makes it challenging to construct and test hypotheses about the shape, something that is achievable with geometric morphometrics such as Procrustes analysis [27]. On the other hand, representations using wavelets or Gabor functions are understudied by botanists, and their interpretation needs to be improved [3]. Moreover, these

approaches need to incorporate expert knowledge, such as botanical manuals or herbariums, which is a significant drawback, resulting in limited adoption of these methods in morphological analysis [26]. This point brings us to the question: *can we employ a contractive representation using base functions that are interpretable by botanists?*

Recently, approaches based on machine learning models aimed to produce categories objectively, which refer to groups of objects with the same visual aspect. For instance, deep clustering is a strategy based on deep neural networks to automatically discover clusters or groups of similar data points in large datasets [23]. This method can identify patterns and relationships in the data that may take time to discover [23]. However, deep clustering is computationally intensive and may require large amounts of data to be trained effectively [84]. On the other hand, novel category discovery, a technique that uses machine learning algorithms to identify new categories or concepts in a dataset [121, 78, 63, 56]. This approach is helpful in cases where the data is highly diverse, or the number of categories needs to be discovered or determined [78]. Nevertheless, in these approaches can be challenging to interpret the generated categories because of the deep learning approaches black-box nature [84].

The main drawback of current approaches for category discovery is that they need to clarify from where the discovered categories come. These approaches generally use all available information without focusing on domain-relevant features [84]. For example, a deep clustering approach may discover categories about all traits in the leaf image dataset rather than a specific trait, e.g., the margin type. As a result, it becomes difficult to interpret the origin of the categories [130]. This lack of interpretability can make it difficult to understand and trust the categories discovered by these methods [79]. Therefore, there is a need for more robust and interpretable approaches to category discovery that can extract relevant features and provide insight into how the categories are formed, especially in life-science areas [8]. In addition, most of these methods lack a formal definition of a category, commonly relying on the ambiguous concept of a group. This way, the categories discovered should be objective, and ideally, their origins should emerge naturally, providing a more robust and interpretable understanding of the data [56].

Justification

Categories are essential for constructing scientific knowledge and understanding reality [64]. Categories refer to groups of entities classified together based on their common properties or features for understanding and evaluating knowledge within a specific field or discipline [8]. The categories are relevant for analyzing and organizing objects, phenomena, and knowledge [8].

The amount of data produced by automatized processes is unprecedented. Therefore, traditional methods for organizing and analyzing these data are required [23]. Categorization plays a crucial role in this process by helping to organize the information and make it more manageable. Additionally, grouping similar items or concepts allows extracting valuable insights and structuring knowledge from data [23]. Furthermore, it allows individuals to place a concept or idea within a larger category and understand its relationship to other similar ideas [8].

Qualitative shape analysis and geometric morphometrics are current approaches used for morphological description, but they have the drawback of heavily relying on manual input and being time-consuming [26]. Although various automated and objective methods have been proposed for morphological analysis, they are not widely used because they diverge from how botanists typically work [3].

On the other hand, recent studies warn that the number of experts in taxonomy and botany, in general, is declining [107]. In an era with increasing data and ecological, climatic, and food-related issues that significantly impact life and sustainability, automating processes to align with botanists' methods can substantially enhance their efficiency and effectiveness [107].

The automation of morphological analysis could represent a significant advancement for modern botany. By reducing the manual workload and time required to analyze plant shapes, experts would have more freedom to focus on creative and investigative tasks. Moreover, streamlining the identification and classification of species through automation could lead to more effective responses to current challenges related to the environment, conservation, and food security.

Objectives

In this research, our primary goal is to develop a computational model capable of automatically and interpretably discovering leaf shape categories from a given set of samples. By harnessing advanced algorithms and techniques, our model can process the input data effectively and extract essential morphological features from the leaf traits. Proposing a representation space that captures leaf traits' inherent variations and diversity is imperative. This objective involves identifying and defining crucial morphological features with high discriminatory power, allowing the model to precisely distinguish between different leaf shapes.

A key challenge in this research is to devise a category discovery strategy, both robust and

objective. The computational model will be designed to autonomously identify visual categories without requiring manual sample data annotation. This approach minimizes human biases and laborious manual efforts, making the categorization process more efficient and reliable.

Transparency and interpretability are fundamental aspects of any computational model in scientific research. Hence, we will meticulously design a strategy for explaining the shape categories discovered by our model in a transparent and interpretable manner. The model will generate quantitative results, such as cohesion and similarity values for each category, and qualitative explanations, including visual representations of data distribution in the morphospace and the formation of groups in the dendrogram. These comprehensive analyses will enable biologists and researchers to effectively understand and validate the categorization results.

By accomplishing these specific objectives, our work aims to significantly contribute to automated morphological analysis in botany, providing researchers with a powerful tool for exploring leaf shape variations in an interpretable and efficient manner. The model's potential for uncovering new insights and discoveries in leaf morphology promises to advance ecological and biological studies and foster collaboration between computational researchers and domain experts.

Main contributions

The main contributions of this computational model for automatically discovering objective, robust, and interpretable visual categories with applications in biological shape analysis are highlighted below.

- **Automated Categorization without Annotations:** The proposed model can discover visual categories without requiring manual sample data annotation. This automation eliminates manual annotations' laborious and time-consuming tasks, freeing biologists to focus on higher-level interpretations and hypothesis testing.
- **Flexibility and Adaptability:** The model's flexibility allows it to adapt to different morphological features and research contexts. Biologists can tailor the model to their needs, enhancing its interpretability and adaptability to their workflow.
- **Quantitative and Qualitative Analysis:** The model produces quantitative results, such as categories with cohesion and similarity values, and qualitative explanations, like data distribution in the morphospace and dendrogram formation. By merging both perspectives, the model provides a comprehensive and insightful analysis.

- **Bridging Objective Data Analysis and Biological Interpretations:** The model bridges the gap between objective data analysis and meaningful biological interpretations by providing quantitative and qualitative outputs. This dual perspective enriches the biologist's understanding of shape variation and facilitates integration into current morphological studies.
- **Exploring Relationships with External Factors:** The model's versatility allows for exploring relationships with external factors such as climate, ecology, soil characteristics, environmental pressures, and crop production. This potential connection can support broader environmental and ecological studies.
- **Handling Unclassified Specimens:** The model handles unclassified specimens, which can lead to significant ecological or biodiversity findings, such as discovering new species, identifying anomalies, and characterizing biodiversity.
- **Formal Definition of Visual Categories:** The study clearly defines visual categories as "a set of groups that persist at different scales and exhibit high cohesion." This definition is supported by various concepts, making the category concept more robust than a simple cluster.
- **Independence of Representation and Category Discovery Components:** The model's representation and category discovery components are independent, allowing it to be applied to multiple problems requiring category discovery by changing the representation component.
- **Additional Computational Methods:** Besides the visual category discovery model, several original computational methods are provided for tasks such as leaf image binarization, segmentation, density adjustment, feature extraction, and data visualization.

Overall, this computational model represents a significant advancement in automating morphological description and offers valuable biological shape analysis tools. By providing automated categorization and comprehensive analysis of morphological traits, the model has the potential to revolutionize ecological and biological studies, leading to valuable discoveries and insights.

Thesis structure

This thesis is organized as follows: The second section establishes the method's baseline for discovering shape categories. It introduces an experimental framework for species identification based on a sample. Additionally, it outlines the workflow for automatic morphological description. Section three details the development of a novel computational model for discovering robust shape categories. The chapter addresses whether it is feasible to develop a

computational model capable of uncovering objective, robust, and interpretable categories of leaf-shape traits. In addition, algorithms are proposed to extract features from specific shape traits. Robust categories are defined based on concepts related to topology and the optimization of neighborhood relationships. In the final section, we present the work's contributions and the most relevant findings and insights in the morphological description area. In addition, we comment on the conclusions and limitations of the work.

2 Contour analysis for interpretable leaf shape category discovery

Abstract

Background The categorical description of leaf shapes is of paramount importance in ecology, taxonomy, and paleobotanical studies. Classification systems proposed by domain experts support these descriptions. Despite the importance of these visual descriptive systems, classifications based on this expert’s knowledge may be ambiguous or limited when representing shapes in unknown scenarios, as expected for biological exploratory domains. This work proposes a novel strategy to automatically discover the shape categories in a set of unlabeled leaves by only using the leaf-shape information. In particular, we overcome the task of discovering shape categories from different plant species for three different biological settings.

Results The proposed method may successfully infer the unknown underlying shape categories with an F-score greater than 92%.

Conclusions The approach also provided high levels of visual interpretability, an essential requirement in the description of biological objects. This method may support the morphological analysis of biological objects in exploratory domains.

2.1 Background

Visual shape description in plants is a very specialized and time-consuming task [59, 122]. Botanists and ecologists require straightforward approaches to communicate relevant information about plant morphology. The construction of category systems allows the communication of the underlying phenomena and the standardization of biological studies [59]. Visual categorization is also an essential task for botanic manual construction, in which expert knowledge is commonly registered as visual categories [7, 60, 35, 11]. In these systems, botanists define key terms accompanied by a visual description of the observed characteristics, with which categories of the shape are established. In systematic biology and taxonomy, experts are extensively trained to perform this task [99, 66].

Leaf categorization based on traditional botanical manuals has several limitations. First,

there are exploratory scenarios in which the working hypothesis is related to datasets of high morphological variability [59, 22], as in poorly explored ecosystems like the Páramos [13]. These scenarios may require particular categorization systems, not necessarily existing, in the commonly used botanical manuals [22]. Second, human-based labeling may be biased by individual opinions because of the high level of subjectivity implicit in the recognition process of biological objects [29]. Finally, botanical manuals are restricted to narrow biological domains. For instance, Northern United States [53], Indian forest [17], tropical Africa [70] and Carolinas in United States [97]. The characterization of unknown biological scenarios cannot necessarily be carried out using this manual [59].

An alternative to characterize plants objectively is digital plant morphology [29]. This approach provides quantitative representations of the object appearance [68, 16, 20]. Several plant science problems have been tackled using this method [12], specifically, species classification and characterization of morphological traits in response to changes in environmental or genetic conditions using, for instance, pseudolandmarks or harmonics to characterize the variation of geometric traits of the leaf contours [68, 16, 20]. However, despite the utility of these approaches to quantify shape, they are limited to object contours with the same homology [71]. Other tools currently available for performing morphometric measurements, like *plantcv* [46], *morpholeaf* [14] or *MowJoe* [37], do not consider automatic approaches to overcome the construction of visual categories systems to describe shapes in the biological domain.

Besides category discovery, the visual description of biological forms also requires high levels of interpretation [77]. Thus the expert should understand the causes of the existence of a shape category, and these categories should also be biologically meaningful [33]. This property of interpretability is fundamental because the knowledge of these causes may help to find explanations of the underlying phenomena, relating the shape to adaptation, function, and development, among other biological features [22]. To achieve these interpretations, biologists commonly use high-level concepts to characterize leaf shape [28]. For instance, the concept of the type of blade or the kind of margin. Notably, these two concepts are closely related to low and high frequencies of the object contour and are captured by the Fourier transform of the border [20]. This fact suggests the use of the Fourier transform representation for recovering some high-level categories used for the foliar description task.

In this work, we propose a novel method to discover the shape categories underlying a set of non-annotated samples based on contour analysis. We show that the use of strategies based on harmonics allows for building a representation space that captures some of the high-level features commonly used by botanists and ecologists in the description of geometrical blade information. We study exploratory scenarios with no known shape categories, in contrast to previous works that focused on the problem of plant species classification [119]. It is

important to note that in the proposed approach the contour information is organized in a morphospace objectively. Therefore, the expert may evidence of the characteristics associated with the biological phenomenon under investigation. We also keep high levels of visual interpretability of the shape information, which is an essential requirement in the characterization of biological data that has been not considered in most of the digital leaf morphology studies [20].

2.2 Methods

Figure 2-1 illustrates the proposed method for constructing interpretable visual categories for a set of images. An image database composed of unlabeled leaves is used as input. The contours of each leaf were extracted by using segmentation and contour extraction algorithms. This information was represented with a complex Fourier transform (CFT), and a set of representative harmonics of the leaf information was selected. Then, a dimensionality reduction method was applied to these harmonics to obtain a three-dimensional morphospace of representation. Finally, an adaptive kernel density estimation method determined the shape categories.

2.2.1 Contour extraction

The input dataset contained natural images with controlled background. These images were represented in a saturation channel because they showed a higher contrast between the leaf lamina and the background. Then, the Otsu method provided a leaf segmentation [93]. A closing morphological operator based on a circular structural element of five pixels of radius removed small holes in the binary image. A tracing algorithm extracted the leaf boundary [51]. This method followed the contour points and returned a two-dimensional vector of vertices. The size of this vector depended on the contour length and the image resolution. In order to have a similar representation among leaves of different sizes, a cubic spline-based interpolation was applied to this array [51]. In particular, $N = 512$ samples uniformly spaced were obtained to represent each contour.

2.2.2 Contour representation

A p -type transformation was used for contour representation, this transform corresponds to a CFT representation of the shape information [112]. Before the CFT, each spatial position of the resampled border (x, y) was represented as a complex value $z = (x, jy)$, with $j = \sqrt{-1}$. The points in the border conform a complex discrete signal $z[n]$, with

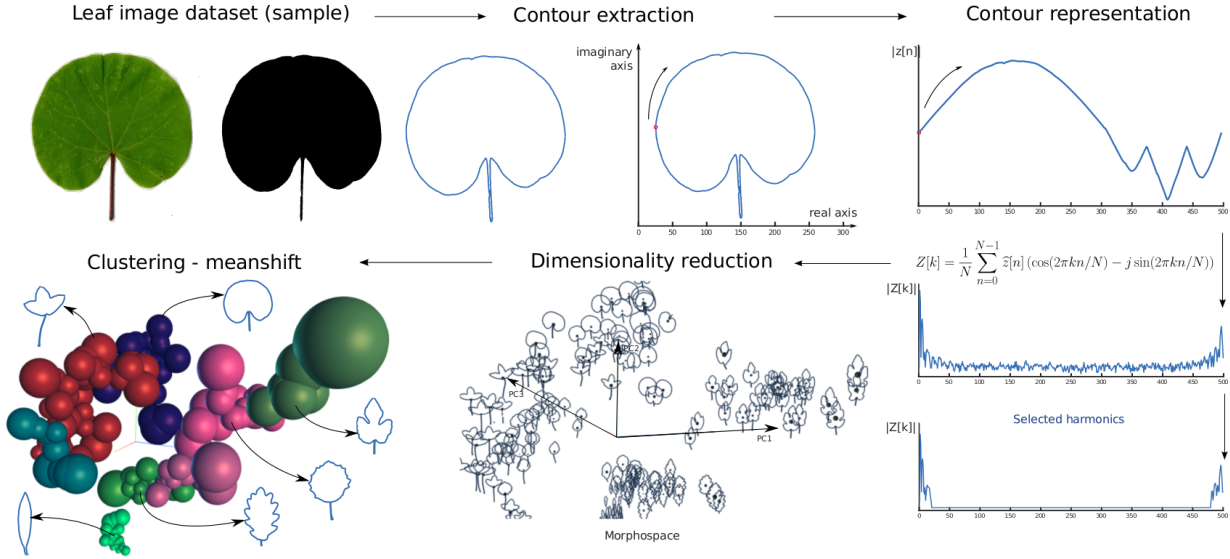


Figure 2-1: Graphical representation of the strategy for category discovery in leaves dataset.

The leaf contours in the dataset were obtained by using binarization and contour extraction. This shape information was represented by a complex Fourier transform. A set of representative harmonics of the leaf information were selected. Following this, a dimensionality reduction method was applied to the selected harmonics. Finally, an adaptive kernel density estimation method was used to determine the shape categories.

$n = 1, 2, \dots, N$. Later, the slopes $\Delta z[n]$ among the adjacent points in $z[n]$ were computed as $\Delta z[n] = (x[n+1] - x[n], j(y[n+1] - y[n]))$. This representation provides robustness to rotation transformations. The slopes were normalized by the distance $\|\Delta z[n]\|$ among the neighbor points n and $n+1$, as follows: $\hat{z}[n] = \frac{\Delta z[n]}{\|\Delta z[n]\|}$, with $n = 1, 2, \dots, N-1$, this normalization provides invariance to scale transformations. Later, a CFT was applied to the normalized slope signal $\hat{z}[n]$, obtaining:

$$Z[k] = \frac{1}{N} \sum_{n=0}^{N-1} \hat{z}[n] e^{-j2\pi kn/N}$$

where k is the harmonic index, $N/2 + 1$ is the maximum frequency order and $Z[k]$ is the k -th harmonic. For the contour description, it is not essential to use the complete set of harmonics [119]. Therefore, the number of harmonics was 22, which allowed a suitable reconstruction of the leaf contour [92].

2.2.3 Dimensionality reduction

Following previous works in the analysis of foliar shapes [29, 115], a dimensionality reduction based on the Principal Component Analysis (PCA) was applied to the selected harmonics.

This process was performed by using a Singular Value Decomposition (SVD) of the covariance matrix computed using the complex harmonics [115]. In order to have visual interpretability of the representation space, the first three principal components were studied.

2.2.4 Clustering

After dimensionality reduction, the category discovery process was performed. For this, the low dimensional data was firstly normalized by applying a whitening transformation in each dimension [34]. A shape category was defined as a cluster emerging in the previously constructed representation space. In this work, two clustering approaches were explored, namely, meanshift [44] and adaptive meanshift [103].

The meanshift algorithm is a non-parametric clustering method for locating the maxima of a density function given n discrete data sampled from that function [44]. Given n data points $u_i, i = 1, \dots, n$ on a d -dimensional space \mathbb{R}^d , the multivariate kernel density with kernel $K(u)$ and bandwidth h parameter is given by:

$$\hat{f} = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{u - u_i}{h}\right).$$

This algorithm provides the modes of the density function, which in our case corresponded to shape categories. The meanshift algorithm directly provides multiple clusters, in contrast to other approaches like k-means which require a definition of the number of classes beforehand. Nevertheless, meanshift results are highly dependent on the bandwidth parameter selection, which indirectly determines the number of classes.

An adaptive meanshift algorithm was also explored to overcome the bandwidth selection issue. This algorithm uses the local density in the representation space to define a dynamic bandwidth for each sample. In particular, the Euclidian distances between u_i and its first k neighbors were averaged and then used as the sample bandwidth parameter h_i [103], i.e., these h_i were used for computation of the meanshift vector. In the proposed setting, the k parameter was obtained experimentally using the TreeMew dataset; six groups of species compose this dataset. The average distance to the eight nearest neighbors allowed the recovery of these six groups in TreeMew; the k parameter was then fixed for all experiments.

2.2.5 Evaluation

Leaf image dataset

The category discovery task consists in arranging a non-annotated dataset in a representative set of shape categories and providing them with coherent explanations in biological terms. There are several public leaf datasets available to study plant species that can be used for evaluation purposes. In this work, two leaf image annotated datasets with information about species with different morphology were selected, namely, TreeMew and ImageClef 2014 datasets [48]. These datasets contain high-quality and quantity isolated leaf image samples, all of them with a controlled background. These conditions helped to extract good-quality contours. Each image in these datasets is annotated with the plant species, which was used as ground truth for the shape category discovery problem. Figure 2-2 shows a sample of each species selected in this study.

For the quantitative evaluation, the samples were organized in three sets to perform shape category identification. The TreeMew was used to build a test set (TreeMew) with six groups, with 20 samples per group. Similarly, for the samples in the ImageClef database, two test sets were constructed (Clef30a and Clef30b), each one with six groups, and 30 samples per group. Table 2-5 shows the corresponding morphological description, which was obtained by using the Hickey manual [35]. As observed, the selected species show differences in their blade shape and margin. It is expected that the proposed method can discriminate samples in different classes using these two criteria. Importantly, these sets have high morphological variability, as Table 2-5 shows. Therefore, this experimental setting is appropriate for evaluating the category discovery strategy.

Experimental settings

The evaluation was two-fold: a quantitative evaluation, to assess the method’s capacity for recovering the original categories, and a qualitative evaluation, to study how the method characterized biologically relevant morphological leaf traits related to the extracted categories.

The shape category discovery problem aims to predict shape categories presented in an unlabeled sample set [55, 129]. We assumed that each plant species corresponds to a different shape category. Under this assumption, the original species of each sample constituted the ground truth for the category discovery problem. A confusion matrix and the corresponding F-score provided quantitative measures of the method performance in the identification of these categories. This last measure considers both the precision and the recall of the class discovery tasks [2]. A leave-one-out scheme was used to study the variability of this performance measurement across different datasets. Once the samples were projected into

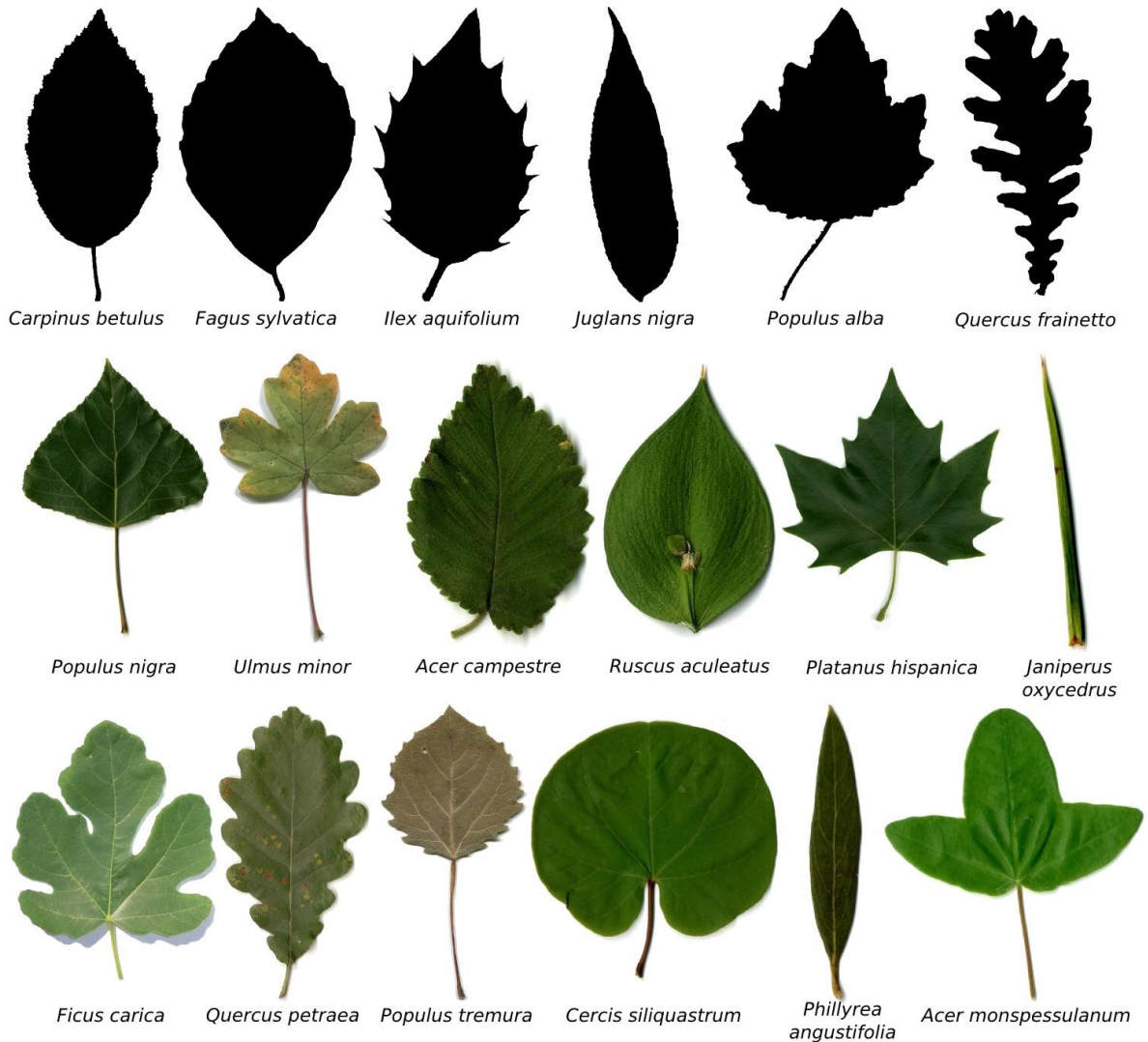


Figure 2-2: Groups of leaves samples used for testing the method. The image shows the selected species from the TreeMew and ImageClef datasets. The species with the most quantity of samples were selected. The leaves groups were organized in the following way, Top: *Tree leaf database MEW 2010*, middle: *Clef30a*, and bottom: *Clef30b*.

the reduced representation space, the clustering algorithm was applied for three different configurations of distance and clustering algorithm, namely:

- Data whitening and meanshift algorithm $MS + W$. Data whitening consists in subtracting the mean and dividing by the deviation of the data in each dimension, similar to the Mahalanobis distance [32].
- Data whitening and adaptive meanshift algorithm $AMS + W$.

- Data without whitening and adaptive meanshift algorithm.

Finally, a leaf sample per category was projected over the principal components to perform the qualitative assessment. The linear combination of harmonics in each principal component was shown and joined with projected samples for interpretation. The aim here was to recover the margin types and blade shapes of the leaf samples.

2.3 Results

2.3.1 Capability of the method to recover the original categories

Figure 2-3 shows the morphospace 3D for the evaluated datasets. Each morphospace shows spheres and representative leaf prototypes. The center of the spheres represents the position of each leaf sample for the evaluated datasets. The sphere radius is given by the adaptive meanshift algorithm. The spheres that were displayed with the same color conformed to the same leaf shape category. The prototypes were the representative sample of each cluster discovered. The leaf prototype corresponds to the closest leaf sample using Euclidean distance to the cluster centroid.

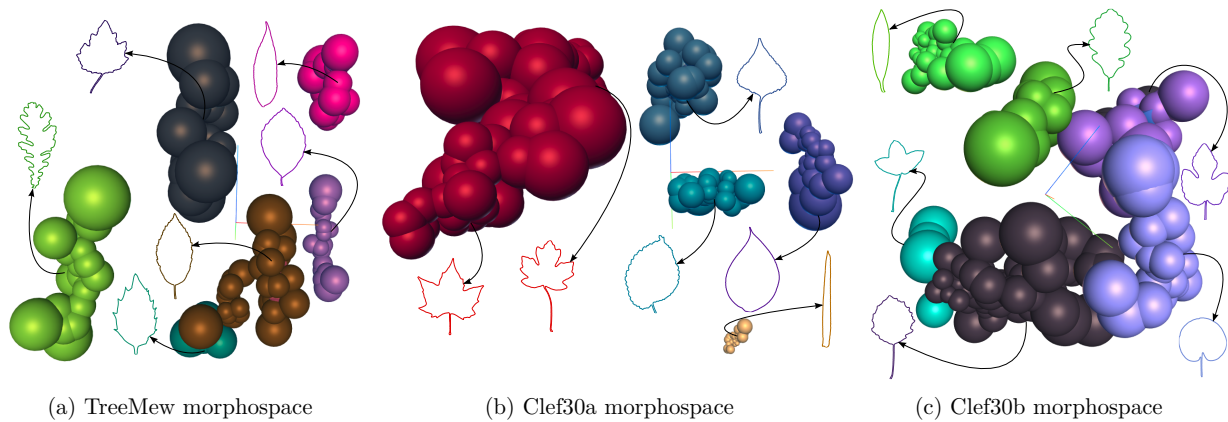


Figure 2-3: Results of adaptive mean shift clustering for the three evaluated datasets. Each morphospace shows spheres and representative leaf prototypes. The center of the spheres represents the position of each leaf sample for the evaluated datasets. The sphere radius is given by the adaptive meanshift algorithm. The spheres that are displayed with the same color compose the same leaf shape category. The prototypes were the representative sample of each cluster discovered.

Table 2-1 reports the quantitative performance obtained by using different experimental settings. In particular, two algorithms: meanshift and adaptive meanshift, and two distances:

Euclidean and Euclidean plus whitening, which is similar to Mahalanobis distance [32]. This was done in the following combinations: meanshift + whitening, adaptive meanshift + non-whitening, and adaptive meanshift + whitening. As observed, the use of adaptive meanshift and whitening resulted in the highest performance for the three explored datasets. High values of F-scores were obtained for the three datasets. Figure 2-3 shows that the proposed representation space locates nearby similar shape samples; additionally, the method was able to separate groups of different species samples. Tables 2-2, 2-3 and 2-4 show the confusion matrix for the evaluated datasets. In the test datasets, the emergence of leaf clusters was evident. Finally, Table 2-2 reports the method recovered most of the ground truth categories associated with the original species.

Table 2-1: Performance comparison between Mean Shift + Whitening (MS+W), Adaptive Mean Shift + Non-Whitening (MS+NW) and Adaptive Mean Shift + Whitening (AMS+W). Table reports the *mean* \pm 1 *SD* for each performance measurement (F-measure).

Dataset	MS+W	AMS+NW	AMS+W
TreeMew	93% \pm 2.1	88% \pm 3.5	97% \pm 1.4
Clef30a	93% \pm 1.4	90% \pm 2.4	97% \pm 1.4
Clef30b	91% \pm 2.3	87% \pm 3.8	92% \pm 2.8

Table 2-2: Confusion matrix results for TreeMew dataset using adaptive meanshift plus whitening. F-measure score 0.95.

Specie group	1	2	3	4	5	6	7	8
<i>Ilex aquifolium</i>	14	0	0	0	0	0	5	1
<i>Fagus sylvatica</i>	0	20	0	0	0	0	0	0
<i>Carpinus betulus</i>	0	0	20	0	0	0	0	0
<i>Juglans nigra</i>	0	0	0	20	0	0	0	0
<i>Populus alba</i>	0	0	0	0	20	0	0	0
<i>Quercus frainetto</i>	0	0	0	0	0	20	0	0

2.3.2 Qualitative evaluation

The proposed method aims also to provide an interpretable representation of the discovered categories. In the experimental setting herein proposed we considered species from six dif-

Table 2-3: Confusion matrix results for Clef30a dataset using adaptive meanshift plus whitening. F-measure score 0.97.

Specie group	1	2	3	4	5
<i>Populus nigra</i>	30	0	0	0	0
<i>Ulmus minor</i>	0	30	0	0	0
<i>Acer campestre</i>	0	0	30	0	0
<i>Platanus hispanica</i>	0	0	0	30	0
<i>Ruscus acuelatus</i>	0	30	0	0	0
<i>Janiperus oxycedrus</i>	0	0	0	0	30

Table 2-4: Confusion matrix results for Clef30b dataset using adaptive meanshift plus whitening. F-measure score 0.92.

Specie group	1	2	3	4	5	6	7
<i>Ficus acrica</i>	24	0	0	0	0	0	6
<i>Quercus petraea</i>	0	30	0	0	0	0	0
<i>Populus tremura</i>	0	0	29	0	0	1	0
<i>Cercis siliquastrum</i>	0	0	4	26	0	0	0
<i>Phillyrea angustifolia</i>	0	0	0	0	30	0	0
<i>Acer monspessulanum</i>	0	0	21	0	0	9	0

ferent shape categories from the TreeMew dataset. Shapes can be described for the complete leaf or their parts as described in Table 2-5. These shape categories were proposed using the Hickey manual [60]. This manual contains high-level shape concepts related to shape, margin, base, and apex. In order to reach high levels of interpretability some leaves were selected from the morphospace to be shown on the representation space axis. For this, we fixed equally spaced points on the axis and the closest sample to these points were shown in the axis, as illustrated in Figure 2-4.

These projections show the morphological variability of the dataset along the main axis. By examining samples in each axis, the shape features that discriminate the groups are identified. As observed, the species with the same margin were closely represented on the first principal component (PC1). Therefore, PC1 represents mainly high-frequency border information that can be linked to these margins. Similarly, the second principal component (PC2) groups species with similar blade shapes, which are projected to the vertical

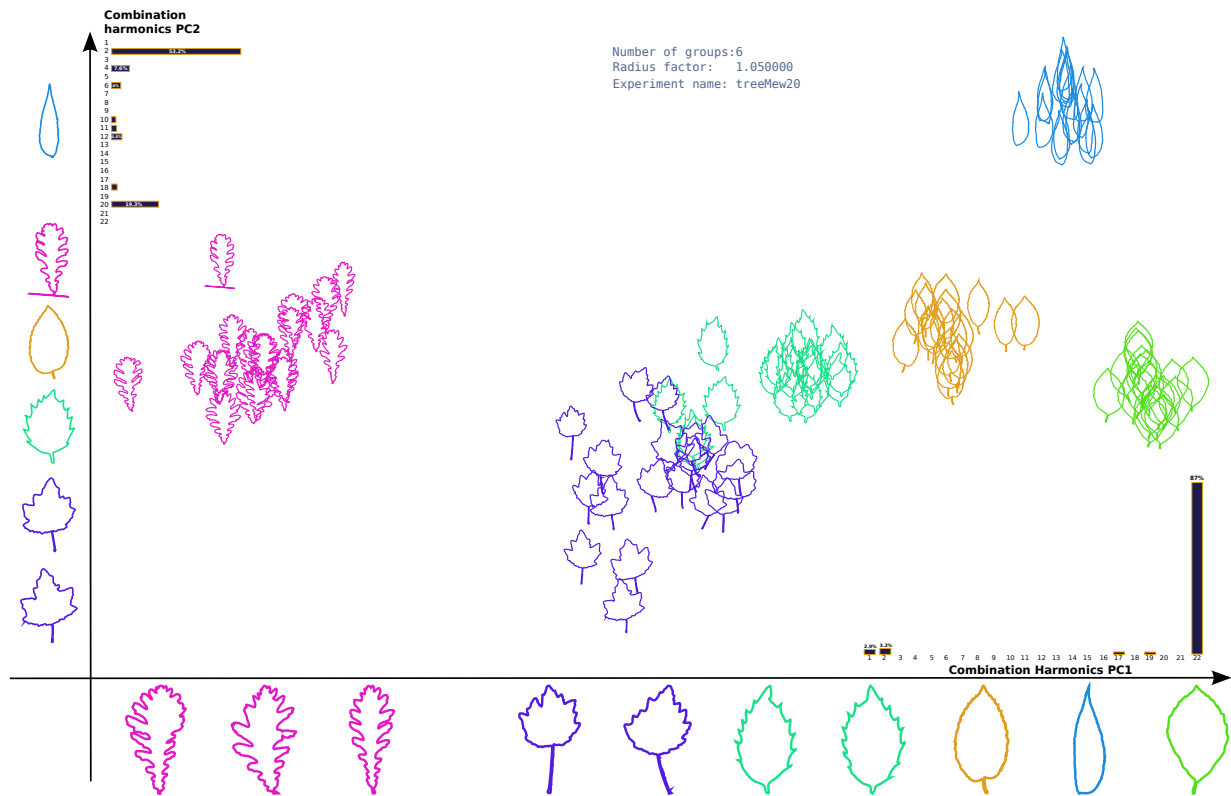


Figure 2-4: Representation space of leaf shape given by PC1 and PC2 for TreeMew dataset. Each axis represents a principal component and shows its harmonics composite. Different contour leaf samples projected from morphospace are shown under the axis. As observed, the x-axis is linked to variations in the margin, while the y-axis is linked to the blade shape.

axis from wide to thin form. More specifically, clusters related to species *Carpinus betulus*, *Fagus sylvatica*, and *Ilex aquifolium* are very close in the representation space, as shown in Figure 2-4. Interestingly, these species also present high levels of similarity according to the botanical manual, as observed in Table 2-5. On the other hand, species *Juglans nigra* and *Quercus frainetto* are far from each other, which can also be observed in the proposed representation space. In the ImageClef dataset, Figures 2-5 and 2-6 showed similar behavior in PC1, corresponding to changes in the margin, while PC2 was related to the leaf width. This result suggests that the method can be used to study margins and shapes simultaneously, resulting in a rich representation.

Table 2-5: Morphological description for the species used in each test group. This description was obtained by using the Hickey manual [60]

Specie	Shape	Margin	Base	Apex
TreeMew:				
1. Carpinus betulus	Elliptic	Dentate	Rounded	Convex
2. Fagus sylvatica	Elliptic	Crenate	Concave	Convex
3. Ilex aquifolium	Elliptic	Serrate	Convex	Acuminate
4. Juglans nigra	Oblong	Entire	Decurrent	Acuminate
5. Populus alba	Ovate	Crenate	Rounded	Convex
6. Quercus frainetto	Obovate	Dentate	Complex	Complex
Clef30a Selection:				
1. Populus nigra	Ovate	Crenate	Convex	Convex
2. Acer campestre	Elliptic	Dentate	Convex	Complex
3. Ulmus minor	Elliptic	Dentate	Complex	Convex
4. Ruscus aculeatus	Elliptic	Entire	Convex	Acuminate
5. Platanus hispanica	Ovate	Serrate	Truncate	Convex
6. Janiperus axycedrus	Special	Entire	Complex	Straight
Clef30b Selection:				
1. Ficus carica	Ovate	Crenate	Crodate	Convex
2. Quercus petraea	Obovate	Dentate	Convex	Convex
3. Populus tremura	Elliptic	Crenate	Convex	Convex
4. Cercis siliquastrum	Elliptic	Entire	Lobate	Rounded
5. Phillyrea angustifolia	Oblong	Entire	Decurrent	Straight
6. Acer monspessulanum	Elliptic	Entire	Cordate	Rounded

2.4 Discussion

A new method for the morphological analysis of leaves is introduced. The method allows the discovery of the categories of leaf shapes in an unlabeled dataset. These categories are interpretable from the biological point of view. The method uses a harmonic representation of the contours, a dimensionality reduction, and an unsupervised clustering strategy. The results show that the strategy identifies categories of leaves related to concepts of margin and foliar lamina. This strategy allows studying sample sets in which the categories are unknown, which may appear in poorly studied biological scenarios.

Results in Table 2-1 show that the proposed approach may uncover the underlying shape categories for different samples of unlabeled leaves, by using only leaf contour information. In particular, the method provided high values of F1-scores (average 95%) in the tasks of discovering previously known shape categories related to the species, by using only unlabeled data. Despite the morphological variability of the datasets herein explored, which includes different kinds of margin, base, and apex, see Figure 2-2 and Table 2-5. The scores and confusion matrices indicate that most of the samples were assigned correctly to the

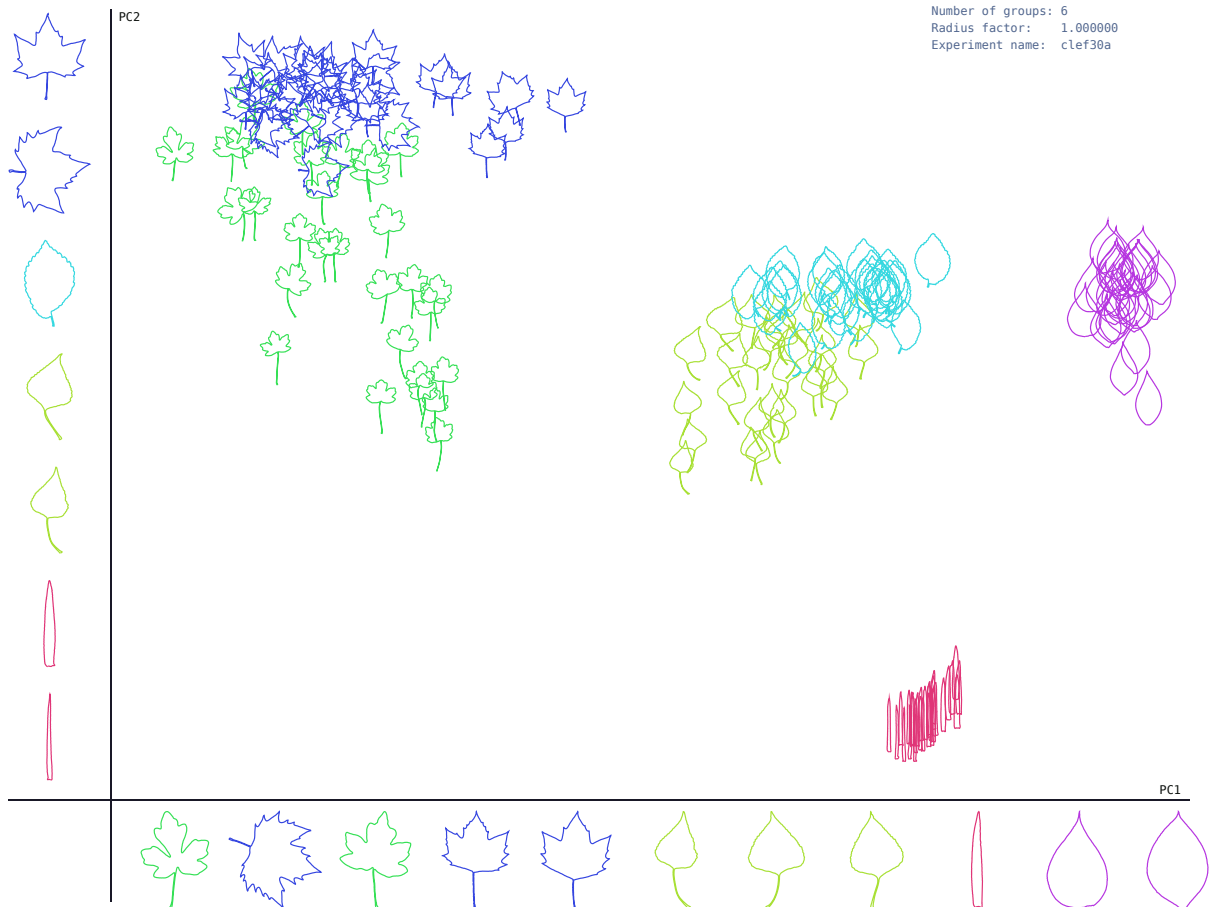


Figure 2-5: Representation space of leaf shape given by PC1 and PC2 for first Clef selection dataset. Each axis shows different leaf samples projected from the morphospace under its principal component.

original shape category. Importantly, no prior knowledge about specific shapes resulted in these categories, in contrast to previous approaches that strongly rely on domain expertise, for instance, particular categories of lamina shapes, as commonly found in botanic manuals [7, 60, 11], or individual landmarks located over the leaf border [71]. Importantly, this expert knowledge may not be available for the description of unknown morphological scenarios [14, 95]. Therefore, the proposed approach is relevant for this kind of description.

In principle, in unknown biological scenarios, shape categories are not known beforehand and may differ from ones used for known scenarios [29]. For instance, in the plant communities in the tropical region, such as Páramos and Guyana Shield [13], for which recent evidence suggests a high morphological variability and endemism. There are no botanical manuals for these scenarios, and the existing ones are not from the region and probably cannot explain the variety of forms. To discover these categories, we used a highly flexible

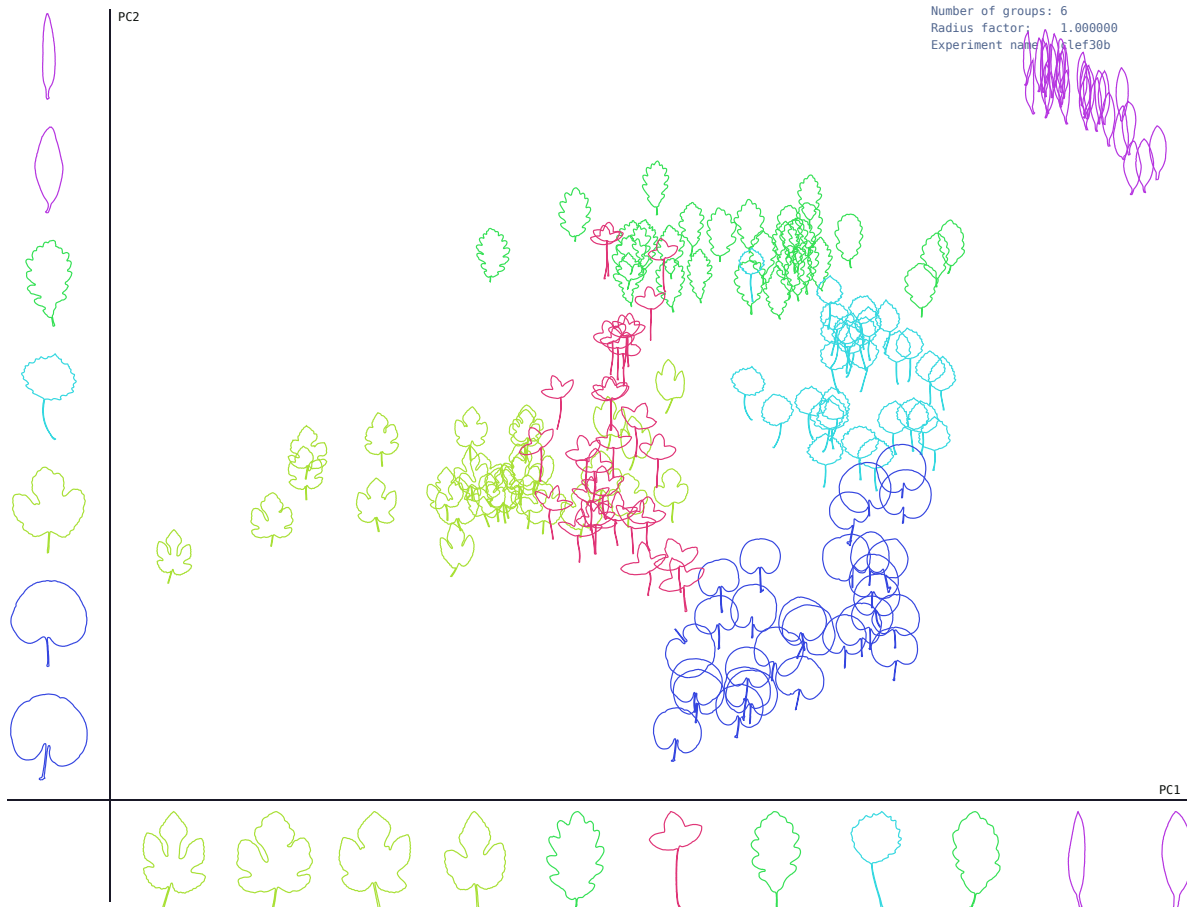


Figure 2-6: Representation space of leaf shape given by PC1 and PC2 for second Clef selection dataset. Each axis show different leaf samples projected from the morphospace under its principal component.

low-level representation space that captures biologically meaningful information about the leaf border, in particular, its large and fine variations [13]. The proposed representation exhaustively captured a broad set of lamina border variations in the Fourier harmonics, providing a rich morphospace to represent possibly unknown sample morphologies. We assumed that leaves with similar variations in the border were close in this morphospace. Therefore, shape categories associated with common morphological features are expected to emerge as clusters. Results in Table 2-2 show that the clusters or shape categories identified in this space, using only the available samples, coincide with the ground truth of shape categories. Remarkably, these categories resemble known shape categories for different classes using only endogenous information from the sample. To our knowledge, this result constitutes the first evidence about the possibility of automatically discovering categories of the shape of biological forms. Alternative approaches have been proposed to discover these categories in natural images [52, 55, 129]; however, these approaches have not been explored yet for the

discovering of leaf shape categories problem.

Low performance observed in F1-scores for some of the studied scenarios is linked to two principal causes. First, a high level of morphological overlap among some of the original shape categories. For instance, in the dataset Clef30a the species *Ulmus minor* and *Ruscus acuelatus* have high levels of visual similarity, see Figure 2-7b, resulting in a single shape category, see category number two in Table 2-3. Despite that, the proposed representation was flexible enough to delimitate both categories properly, see curves in Figure 2-7b. Importantly, the visualization considered in the model helps to localize and correct errors in the final assignment of the sample category. Second, in some cases, leaf border information was not adequately represented by the Fourier transform. For instance, this representation did not correctly capture border information for samples in specie *Populus tremura* in violet color in Figure 2-7a, probably because of the presence of high-frequency information in the serrations [15, 50]. Further investigations may also consider alternative data representations which account for these shape particularities [15, 50, 72].

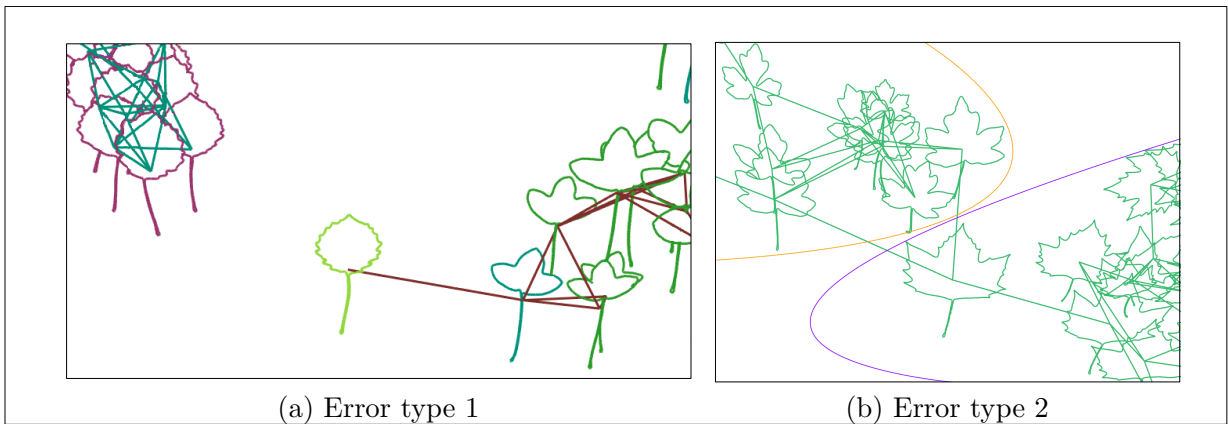


Figure 2-7: Errors examples, in our approach, leaves with similar shapes form clusters. The lines that appear from the leaf center show how these leaves are connected to shape groups. In the left box (a), there are two groups, and in the right box (b), there is one. However, on the left, there is a light green leaf that has a shape similar to one group, but it is connected to another. In contrast, in the right box, all the leaves are connected to the same group, but these could be separated into two species by the violet and orange lines.

Figures 2-4, 2-5 and 2-6 show that common high-level concepts with biological meaning emerged from the representation spaces obtained by PCA projections. Particularly, in the PC1 axis, serrations change from left to right, from serrate margin to entire margin. While in the PC2 axis, leaf shape changes from bottom to top, from wide lamina to narrow lamina. Therefore, we conclude that the major axis relates to the lamina shape concept and

the minor axis to the border serration. These high-level concepts represent explanations for the shape categories discovered [60, 35, 11]. Explainability refers to a human-interpretable description by which the method categorized a shape given a set of unlabeled images [33]. In this case, each discovered category is characterized by a particular combination of lamina and margin shapes. The lamina and serration shape concepts are commonly used by experts to explain leaf shape categories, and they are fundamental for interpreting biological variations [60, 35, 11]. These two factors are at the base of leaf descriptive systems of shape categories, and they are commonly used for taxonomical classification, and leaf adaptation to environmental conditions, among others.

In order to identify the possible factors associated with the obtained shape categories we performed a posthoc analysis to identify. This kind of analysis is also used in other approaches, for instance, Procrustes and Fourier analysis [71], which consider a subsequent interpretation step aimed to identify sources of variations [71]. In these analyses, experts assign a meaning to observed experimental variations. For instance, correlating shape features with known domain variables. Following a similar approach, we conclude that the shape discovering method provides consistent explanations in biological terms, shape, and margin, to the categories discovered. Future work may consider the automatic identification of the concepts that determine the categories and not only rely on the interpreters' opinions. It is worth noting that previous approaches to category discovery do not consider the issue of construction of biological explanations to support biological interpretations [55, 129].

In this work, a complex Fourier-based representation supported the feature description stage. This transformation provides high levels of visual interpretability [115]. In our experiments, the contours become invariant to geometric transformations, and they were also normalized and centered, as in the Procrustes analysis, but without requiring any landmark. Unlike other approaches to contour analysis, harmonics capture contour variability in the frequency space. Therefore, our approach may serve as a tool to analyze this variability in leaves with a different structure. For instance, the approach can be useful when the contours present different lobular compositions, or in sessile leaves, which do not have petioles resulting in open contours. This kind of description is essential also for the description of poorly studied vegetation, as it happens in the high mountain Andean vegetation [13]. A three-dimensional space obtained by PCA embedded the contour representation, and a non-supervised clustering algorithm was used on this representation space to infer the corresponding shape categories. The aim here was to reduce the dimensionality of the data in 3D space and provide visualizations and interactivity with the samples in the representation space. As Figures 2-4, 2-5 and 2-6 show, the leaves were distributed along with the representation space forming dense groups. The distance between a pair of samples was related to their similarity, and the direction between them revealed the particular feature that differentiates them. When the biologist organizes the obtained sample in leaf categories [20, 76]. This

representation allows both a visual representation of the shape information and a suitable space to solve the category discovery problem.

The present work has some limitations. First, the proposed method uses only endogenous information on the leaf contour morphology to project the sample into a morphospace, suitable to discover the shape categories. Future work may consider including additional information related to the scientific question (for instance, precipitation), which helps explain the sample’s morphological variability. This complementary information can be included, for example, as an additional part of the feature vector that characterizes each sample. Future work may also consider interaction with the experts to construct the richest morphospace, enabling post-hoc verification and modification of the proposed categories according to the expert knowledge. Second, increasing the number of categories may difficult the capacity of the method to discriminate the underlying shape groups correctly. As illustrated in Additional file 1: Figure S3, when considering a reduced number of PCs, the shape categories cannot be adequately discriminated, and more PCs should be included. Therefore, the inclusion of additional PCs should be considered when the complexity of the database increases.

2.5 Conclusions

In this work, we proposed a novel method to automatically discover shape categories from the digital image of leaf samples by keeping high levels of visual interpretability of the shape information. The method is based on a complex Fourier representation of the contour, which is embedded in low-dimensional representation space. An adaptive clustering method with whitening was used to discover the shape categories. The method was evaluated through the task of predicting the shape categories associated with different plant species. Our results suggest that the proposed method successfully discovers the plant categories by using only leaf shape information providing high levels of visual interpretability.

3 Robust visual category system of specific leaf shape traits

Abstract

Describing the leaf's shape is critical for taxonomy, plant function understanding, and unveiling the vegetation adaptation mechanisms to environmental changes. Nevertheless, the shape description of biological objects is challenging as it encodes multiple attributes from genetics and the environment. Therefore, most leaf-shape characterization still depends on experts' knowledge. Significantly, these descriptions rely on visual category systems established by experts in botanical manuals. These systems of visual categories group regularities in shapes that humans perceive to explain leaf traits. These knowledge organization systems are highly informative. However, they also have ambiguity and bias risks. This paper proposes a novel approach for automatically discovering robust visual categories for different leaf traits. The proposed strategy relies on morphospaces suitable for representing shape features shared across particular leaf traits and a novel Topological Data Analysis based algorithm for identifying robust groups of shapes in these morphospaces. Results suggest that this approach automatically recovers visual categorical systems for six leaf traits, which highly resemble those determined by experts in classical botanic manuals and visual categories associated with different species. Moreover, the strategy exhibits biological interpretability, enhancing its value in botanical research. This approach represents a first step toward a quantitative description of morphological variability from the visual categorical systems perspective.

Keywords: Novel category discovery, Unsupervised categorization, Leaf shape, Contour analysis, morphological, Image processing, Topological analysis, Image classification

3.1 Introduction

Plant morphological traits and their relationship to the environment are essential in ecology [90]. For instance, the shape of leaves directly influences the plant's function, and it

represents a robust indicator of its adaptation to the environment [91]. However, shape in biology is complex as it encodes multiple traits resulting from genetics, environment, development, and adaptation [80]. Likewise, the shape is an abstract concept that can be differently perceived depending on the biological context and the expert’s perspective [90]. Recently, biologists have warned that the available approaches for shape description cannot quantify all the information a biological organism provides through its shape [3]. Although there are several quantitative approaches for addressing the problem of shape description [29, 20, 116], there is still a need for methods that exhaustively and effectively describe the abundance of shapes exhibited by nature [3].

Traditionally, biologists have employed a qualitative approach for classifying shapes, often referencing visual category systems from botanical manuals [3]. These manuals serve as comprehensive resources, offering detailed descriptions, illustrations, and clear definitions of botanical terms. They facilitate understanding and identification of various plant species by visually representing plant structures, functions, and classifications [60]. For this reason, they are extensively used in qualitative studies analyzing morphological variability in species due to environmental factors by providing descriptors like color, size, texture, and shape [62, 80, 10, 120, 88]. Typically, botanical manuals, like the Hickey manual (a botanic reference widely used to describe different leaf traits), structure knowledge around shared geometrical features, relying on the concept of shape categories [60].

The ability to parse varying stimuli into discrete categories is a fundamental property of human and animal behavior [60, 41]. The category notion represents a general simplified abstraction based on the regularities perceived in objects [67]. The category concept simplifies the analysis process of highly heterogeneous phenomena [98, 41], as in the case of biological object description. Therefore, categories constitute the foundation of many knowledge organization systems [64]. However, in the case of biology, it is only possible to have some of the potential categories because of the high complexity and variability of the biological objects of interest that prevents a global, totalizing, and complete analysis of them [8]. Therefore, studies based on visual category systems have ambiguity and bias risks [31]. For instance, in the Hickey manual [60], an expert differentiates the *rounded* class from the *convex* class because, in the first one, a “smooth arc” is discerned. This shape assessment is highly subjective and may differ depending on expertise.

Alternatively, quantitative approaches can be used for shape description. These approaches, for instance, describe the leaf shape as a sequence of points on the lamina contour and project this silhouette to a so-called morphospace through a geometrical analysis method [116, 21]. The morphospace provides a quantitative representation of the possible form, shape, or structure of an organism [3]. Furthermore, a distance measure can be defined for this morphospace, allowing comparisons between objects to classify, categorize, or relate the results to the quantitative description of genetic, evolutionary, and environmental features [3]. Nevertheless, unlike botanical manuals, the quantitative methods for shape description focus on the most general shape features, having the risk of missing additional information provided

by experts [25]. In addition, these descriptions need more interpretability and explainability [81]. For instance, these shape description strategies do not account for categorical systems for grouping objects according to shared geometrical features the experts may understand [86, 87].

Recently, deep learning models have also been applied to discover underlying categories in data [114]. These works focus on directly learning a morphospace from a large set of samples and organizing objects into categories [23]. However, in these approaches, the control over the morphospace is minimal because of the non-linear back-box nature of the deep learning architectures [130]. In addition, in these techniques, it is difficult to interpret which traits resulted in the discrimination of the categories, a critical interpretability requirement in the biological context [79]. Other strategies based on clustering [116] capture groups of objects. But groups of samples proposed as categories by these methods need more robustness to the expected shape variations exhibited by biological entities and also to the outliers occurrence [116]. Therefore, a quantitative, automatic, and flexible approach is required to describe the shape in different contexts, allowing the organization of biological entities into categories, as is the case with botanical manuals and field guides. But at the same time, these categories must be objective and reproducible enough to explain the shape of any sample.

This work proposes a novel strategy for automatically discovering shape categories. In contrast to previous approaches for shape description, this approach allows quantitative and qualitative shape descriptions for different leaf traits and considers robustness issues. The proposed system relies on two main components. The first aims to represent the geometrical information of various leaf traits in suitable morphospaces to describe relevant shared geometrical features. The second component constructs a high-dimensional discrete combinatorial structure of similarities between samples. This structure codifies relationships in multiple neighborhood scales among samples in the morphospace, from which robust categories emerge. Compared with related methods, the proposed approach provides a high degree of interpretability of the results in morphological terms through a dendrogram that reveals the dynamics of the formation of visual categories and quantifies the similarity relationships at the level of samples, groups, and visual categories. Results suggest that the proposed approach is highly effective in automatically discovering visual categories of leaf shapes when comparing them to the ones defined by experts, using only endogenous shape information and no annotated data. In addition, results show that it is possible to establish a taxonomy of shape groups according to the highlighted shape feature descriptions. Remarkably, the results extended up to six different types of leaf traits. These results suggest that many studies of morphological variability can be automatized, helping to advance in taxonomy, systematics, recognition of new species, and morphological changes caused by global warming [80, 10, 120, 88]. Furthermore, the proposed methodology for discovering categories may apply to other areas of knowledge.

3.2 Related Work

Two approaches are the most widely used in the quantitative description of biological shapes: the geometric approach and the analysis using contour basis functions [3]. Furthermore, techniques for dimensionality reduction are also usually applied after the representation stage, followed by clustering methods to group the samples [29].

The geometric approach uses a set of landmarks to define the shapes to be analyzed and looks for the geometric transformation that minimizes their differences [1]. This strategy has the advantage of being strongly linked to the research question and therefore provides a high level of interpretability of the results [127]. For instance, geometric approaches allowed to identify genetic markers associated with specific shape characteristics such as serration or lobulation based on Procrustes analysis [1, 106, 43], as well as studying the relationship between genetic variability and morphological variability [131, 54, 19, 26, 27]. However, this approach has significant expert intervention and requires a high level of precision in the location of the landmarks. In addition, in some studies, it is not possible to replicate the complete set of reference landmarks in the whole sample [3], limiting the study scope. Furthermore, the level of specialization in the landmark definition and location also constraints the analysis scope [3].

Quantitative and automatic methods, which use a family of basis functions to construct the morphospace, can operate on extensive collections of leaf samples. The most commonly used approach for analyzing various leaf shape features is the elliptical Fourier transform [68]. Researchers have used the elliptical Fourier for taxonomic or systematic studies [30, 94] and the automatic identification of genus or species [126, 85]. However, this representation is limited to closed contours and does not provide information about the geometric correspondence between points. Alternative approaches use more elaborate quantitative representations or machine learning, followed by a clustering technique to classify or categorize the objects [73, 3, 74, 4]. However, using more complex representation methods may restrict the interpretability of the results.

In general, the category discovery problem has been explored in two ways. The first approach aims to discover visual categories for automatically classifying objects in natural images without annotations [69, 38]. The second, more recent, use deep clustering to generate a learned representation space from the samples. This last approach defines categories using semi-supervised learning, in particular, an auto-encoder architecture and a simple clustering method such as k-means [63, 78, 121, 56]. However, while the learned morphospace obtained by this method may produce good results in the grouping phase, it is difficult to identify which traits of the shape establish the groups. Previous work proposed the adaptive mean-shift algorithm in the morphospace for discovering the underlying categories of an unlabeled subset of any samples [116]. However, the method is susceptible to outliers and may fail with inter-cluster samples. Moreover, the categories resulting from these methods need to be more robust for noisy experimental scenarios.

In summary, the detailed shape description performed by experts is mainly performed qualitatively by comparing samples with graphical models defined in the manuals, commonly organized as visual categorical systems. In addition, there are quantitative approaches for describing the shape. However, these approaches do not discriminate the shape features to generate the categories or must be more robust to face the biological variability expected from field samplings. Therefore, a quantitative approach is required to build robust visual categories for different shape traits.

3.3 Materials and methods

Fig 3-1. shows the category discovery process. The method starts with a dataset of leaf images without annotations and generates a set of robust interpretable visual shape categories representative of the sample for different leaf traits.

The process begins by binarizing the image and segmenting the leaf. Then, an algorithm for extracting the petiole also helps characterize different leaf parts, for example, the apex, the base, and the leaf body. The leaf contour is then extracted and interpolated for different contour sizes. The expert may decide which leaf traits will be used to find the robust categories. Then, particular shape representations are applied depending on the leaf trait to be characterized. These shape features are projected to the corresponding morphospace. Following this, the Principal Components Analysis (PCA) method reduces this morphospace to three dimensions, improving the interpretability of the categories to be discovered. Then a simplicial complex codifying neighborhood relationship of shape among samples is built in the morphospace. Connected components of this simplicial complex are filtered out to characterize highly cohesive groups of samples. Next, the groups that most persist across different neighborhood scales are determined and defined as robust categories. Finally, a dendrogram codifies the dynamic of the discovery category process from the sample.

3.3.1 Data and reference visual categories

There are several public leaf datasets available to study plant morphology, which can be used for exploring the shape category discovery problem. In particular, we selected the ImageCLEF2012 [48] and the TreeMew datasets to study the problem and compare performance with a baseline category discovery method [116].

The ImageCLEF 2012 dataset has been widely used in leaf classification experiments and has the appropriate conditions to evaluate the performance of the proposed method. This dataset provides a good-quality image collection of developed leaves (adults) with one specimen per image. Additionally, the dataset has sufficient images per species to allow multiple experiments. In particular, the dataset contains 11,572 images organized into 115 species, of which 57% are scanned, 24% photos with controlled backgrounds, and the remaining 19% are photos in natural environments. We selected up to 49 species with 50 or more specimens

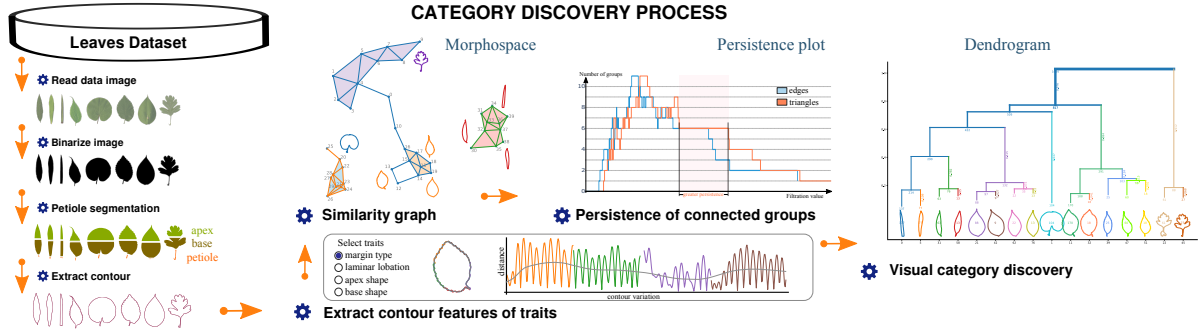


Figure 3-1: Process for the discovery of leaf shape categories. The process starts with images from an unannotated leaves dataset. The first step extracts the silhouette from the image resulting in a binary representation of the leaf. Then, an algorithm segments the petiole, dividing the silhouette into apex, base, and petiole. Once the expert selects a part of the leaf, the method extracts the contour, i.e., the leaf trait for which categories should be discovered. The next step is to extract specific traits to generate the morphospace. Then, a combinatorial structure (similarity graph) codifies sample neighborhood relationships. Finally, the categories returned by the method correspond to the groups with the highest persistence when the neighborhood changes, i.e., the groups that do not change for different neighborhood sizes. Finally, a dendrogram shows the method’s dynamic.

in this dataset for the experiments. In addition, the images are annotated with taxonomy (genus and species), spatial location (latitude, longitude, altitude), locality, and date.

The proposed method aims to discover a set of shape categories in unlabeled samples. To provide quantitative results about the expected performance in the category discovery task, we performed different experiments to identify the categories emerging on samples from related leaf traits labeled in predefined shape categories. The method only considered the sample leaf trait shape and aimed to discover the underlying shape categories.

Therefore, in addition to selecting the data for conducting leaf species discovery experiments, setting a reference system of shape categories is also necessary. For this, the Hickey manual was used [60]. This manual offers a comprehensive and detailed analysis of leaf contour morphology. The manual was meticulously created by esteemed botanists with extensive expertise in the systematics field, aiming to establish unambiguous and standardized terminology for describing leaf forms. The manual provides categorized terms, visual examples, and instructions for accurately describing contour characteristics. These manual reference categories serve as a reference to evaluate the method’s effectiveness in distinguishing samples into distinct shape categories.

In the first set of experiments, four categorical systems of shape defined in the Hickey manual [35], specifically, apex shape, laminar shape, margin type, and base leaf traits, were

taken. Fig. 3-2 shows the visual categorical systems used as reference. Next, following the morphological definitions in the Hickey manual, several samples that accomplish the description of each of the classes in Fig. 3-2 were selected and annotated.

Hickey categories

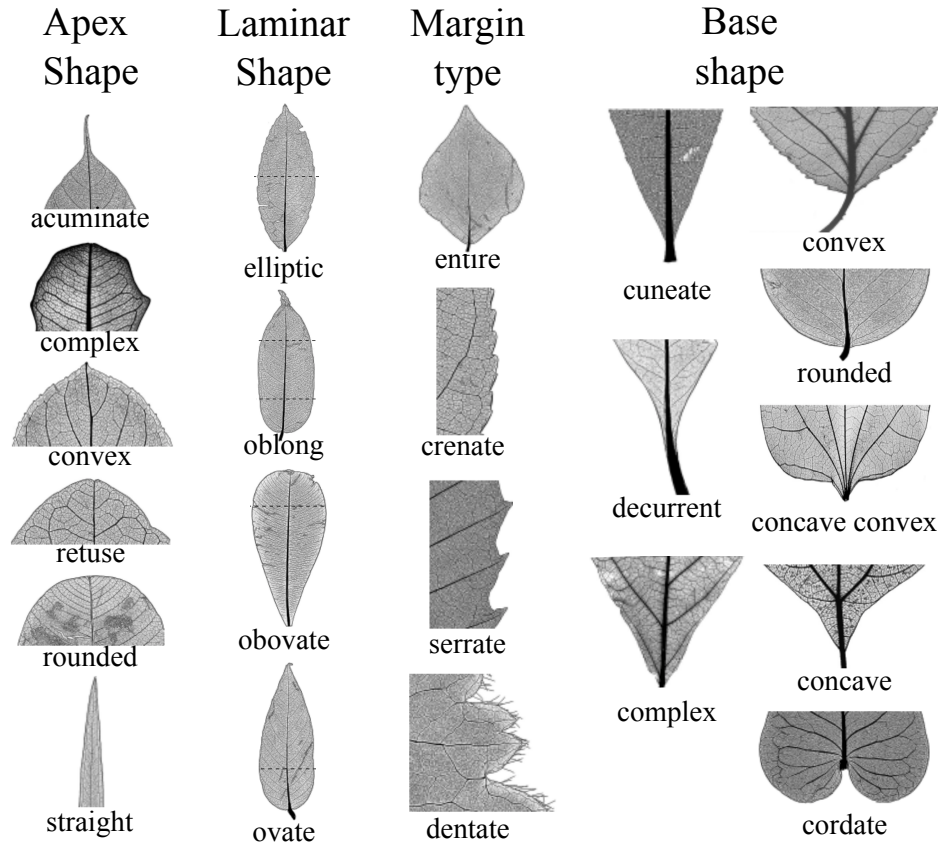


Figure 3-2: Visual categories taken from the Hickey manual [60] for category discovery experiments. In columns, the different shape features were selected for contour description. Composite figure with images taken directly from the Hickey manual [60].

The second set of experiments assessed the method's ability to discover the species of the leaves from the morphological characteristics expressed by the shape. In these experiments, it was assumed that the species were in the same categories. The full TreeMew dataset of 120 samples was used in the first experiment, and ImageCLEF 2012 with eight unbalanced categories and 905 images were used in the second experiment. These annotated categories were considered as ground truth for quantitative analysis.

3.3.2 Leaf segmentation

The first step in the visual category discovery process is segmenting the leaf for each sample in the dataset. The process of segmentation of the leaf and subsequent binarization of the image is the same as that proposed in [116]. This method returns the binary image with the silhouette of the leaf.

3.3.3 Petiole segmentation

Segmenting the leaf's petiole is essential for extracting specific traits from the shape, particularly for segmenting the parts of the leaf, such as the apex, base, or leaf body. In addition, the leaf body helps analyze the leaf's general laminar shape, symmetry, margin, and lobation. Before segmenting the petiole, the leaf is vertically aligned. More specifically, with the help of PCA, the leaf midrib is aligned with the vertical axis. Once the leaf is aligned, the cumulated profile, defined as the number of points that conform to the leaf's interior when intersecting with vertical lines, is computed, i.e., a sinogram of the leaf's interior is calculated [57], as illustrated by the blue curve in Fig. 3-3. To remove the high-frequency variations in the profile, for instance, linked to serrations, the profile is normalized to the zero-one interval, and a low pass filter is applied to smooth the curve, see the orange curve in Fig. 3-3. Then, a gradient is calculated to detect sharp changes in the profile, which may characterize different leaf traits, as observed in the green curve in Fig. 3-3. For identifying the petiole, the method marks the initial points of the first (init point line) and the second (inflection point line) positive slopes of the gradient curve, from left to right. These two points define the segment containing the petiole. The apex (apex point) is indicated by the initial point of the first slope, from right to left. The midpoint corresponds to the middle between the inflection and the apex points.

3.3.4 Contour extraction

The contour extraction process aims to represent the leaf shape information as a vector of two-dimensional points. These points are equidistant vertices from the leaf outline. The number of vertices was set to 640. The Susuki's algorithm extracted the leaf contour [109]. This algorithm takes the binary image, follows the border between the region and the background, and registers the pixel coordinates in an array. Finally, an interpolation algorithm allowed the representation of the leaf contour with the same number of points. For this, the algorithm constructs an equally spaced partition of the curve length to locate the vertices of the contour.

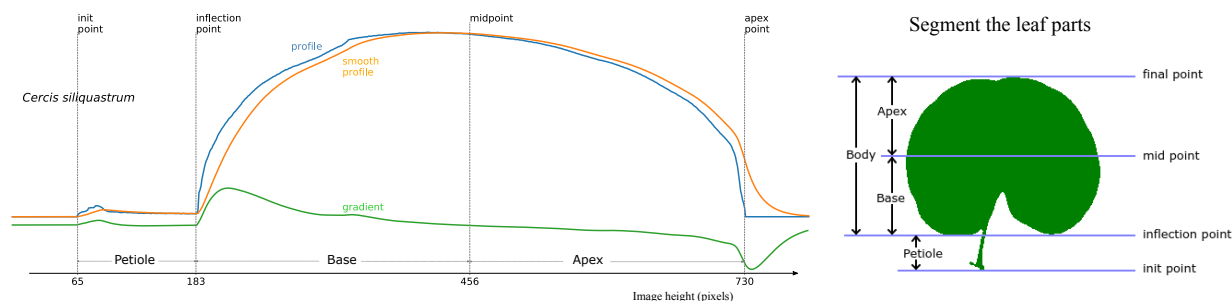


Figure 3-3: Method for segmenting the petiole. This strategy aimed to identify the divisions that separate the leaf into three sections (right panel): petiole, base, and apex, as shown in the right panel. On the left, the panel illustrates the petiole segmentation process. First, the leaf is positioned horizontally, with the petiole at the left. The cumulated profile (blue curve) corresponds with the number of points that conform to the leaf’s interior when intersecting with vertical lines. The orange curve corresponds to a smoothed version of the leaf profile. The green curve corresponds to the derivative of this smoothed profile. For identifying the petiole, the method marks the initial points of the first (init point line) and the second (inflection point line) positive slopes, from left to right. These two points define the segment containing the petiole. The apex point is indicated by the initial point of the first slope, from right to left. The midpoint corresponds to the middle between the inflection and the apex points.

3.3.5 Feature extraction

After contour extraction, particular representations were proposed to describe different parts of the leaf, for which there is an interest in determining shape categories. These representations are similar to the ones commonly used in geometric morphometry [131, 94].

Ideally, to obtain categories that explain the underlying phenomenon in biological terms, the morphospace must satisfy two properties [1]. First, samples with similar shapes should appear close by, and samples with different shapes should appear far away in the morphospace. Second, the similarity relation in the morphospace must be maintained invariant to scale, rotation, translation, and other geometric transformations of the samples [104]. Another desirable property is that the density of the samples in the groups shows slight variation across the morphospace. When a feature space satisfies these properties, it is referred as contractive [104].

In research problems related to the shape of the leaves, shape categories can be obtained from various parts leaf traits, such as the base, the apex, the petiole, the entire leaf, or the leaf without the petiole. Complementary, specific shape features can be extracted for symmetry categories, margin types, and lobation. In each case, specific representations to describe these leaf traits are required to extract the representative features. The algorithms for computing these representations receive a set of contours as two-dimensional arrays of points and

construct particular morphospaces. The algorithms for computing specific representations for different leaf parts are described below.

Shape representation

The proposed method categorizes the shape into closed (whole leaf) and open contours (base, apex, or petiole). The feature extraction method uses a family of p-type functions. The representation method is detailed in [116]. The first 50 p-type coefficients are selected and then reduced to three dimensions using PCA.

Representation of the laminar shape

The representation selected to describe the laminar shape should be powerful enough to capture the different shapes, for instance, *ovate*, *obovate*, *oblong*, and *elliptic* among others [60], but should also be able to discriminate among these different categories.

The algorithm allows representing the laminar shape features, for instance, the classes in the Hickey manual *ovate*, *obovate*, *oblong*, and *elliptic*, among others. For this, a vertical profile of the leaf is built. This profile consists of a series of values indicating the leaf's width at each vertical point. The series is analyzed with the wavelet transform [128], and 16 scale coefficients covering the entire profile are selected. Then, PCA is used to reduce from 16 values to a three-dimensional morphospace.

Characterization of the leaf-lobes

This method is designed to represent the leaf-lobes features present in the samples. The algorithm also uses a family of p-type functions to analyze the contour. Two contours are reconstructed, one with the first six p-type complex coefficients and the other with the first 16 complex coefficients. Then, the point-to-point Euclidean distance between the two contours is calculated, and a one-dimensional signal is obtained. The signal is also analyzed with the wavelet transform [128], and gain 16 scale coefficients are selected that cover the entire signal. Finally, PCA is used to project features to a three-dimensional morphospace.

Characterization of the margin types

The proposed algorithm is designed to represent the kinds of margins, such as *dentate*, *serrate*, *cuneate*, and *entire*, among other shape categories. As before, the process begins by representing the contour using a family of p-type functions. A contour is reconstructed with the first 16 complex p-type coefficients (see Fig. 3-4 columns Harmonic(16)). Then, the point-to-point Euclidean distance between the original and Harmonic(16) contours is calculated, resulting in a one-dimensional signal (see column Shape distance to original in

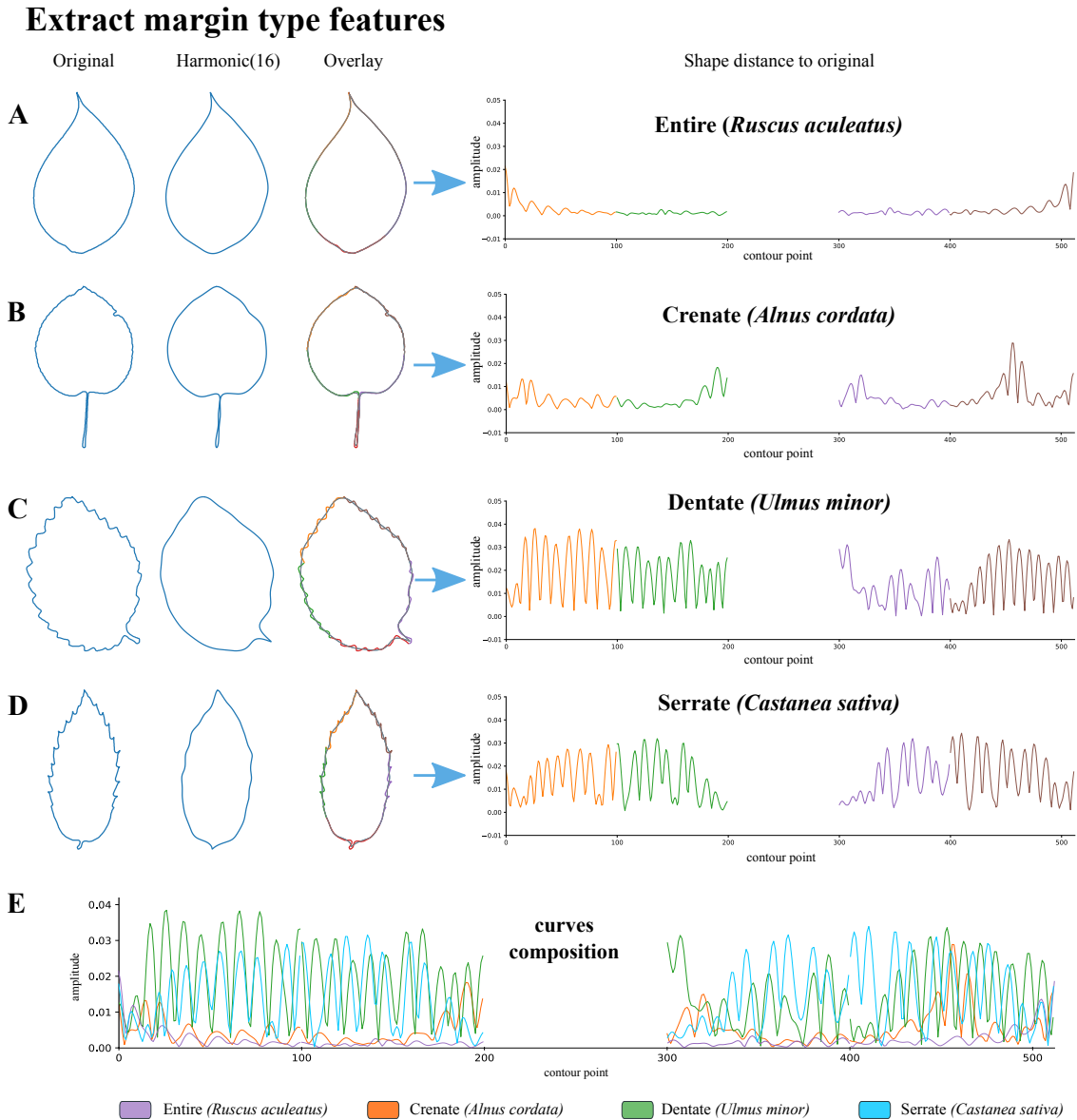


Figure 3-4: Extraction of margin-type features from the leaf contour. Each row represents an observed Hickey category of margin type: A. Entire, B. Crenate, C. Dentate, and D. Serrate. The sequence from left to right includes the original contour, the reconstructed contour with 16 harmonics, the superposition of both contours and a plot illustrating the point-to-point distance between them. The original contour is depicted using a pattern of four colors, indicating different contour parts, starting from the top contour point. In the distance plot, the segment in the center corresponding to the petiole was removed. Part E shows a composition of the distance curves for each category, note how each category exhibits a distinct amplitude and distribution.

Fig. 3-4). This signal is analyzed with the wavelet transform [128], and the largest 32 high-frequency coefficients that characterize the margin variations are selected. Finally, PCA is used again to reduce from 32 values to a three-dimensional morphospace.

3.3.6 Category discovery algorithm

The proposed algorithm synthesizes the information of samples in the morphospace as a set of robust categories. For this, the leaf trait samples are represented by N points in a morphospace of dimension d , with d the number of features describing each leaf trait, in this case $d = 3$. Following Topological Data Analysis (TDA) [123, 24] strategies, the category discovery algorithm constructs N balls of varying radii around each sample, as illustrated in panel A at Fig. 3-5 to build robust and cohesive neighborhoods of similar shape, which are defined as categories.

Defining shape neighborhoods

Each ball represents a *shape neighborhood*, i.e., the regions in the morphospace with similar shapes to the sample in the ball center. Given a set radius (one per ball), two, three, or more balls may eventually intersect. For instance, see the intersections of two balls (2 and 3) or among three balls (15, 16, 17) in panel A at Fig. 3-5. When the intersection among a set of balls is not empty (see intersection of balls 2 and 3), there is a shared shape region in the morphospace, indicating that the corresponding samples are similar in shape.

Complementary, the union of intersecting balls represents a region in the morphospace where shapes are similar, i.e., the union of these balls forms a new neighborhood. For instance, the union of balls centered at 2 and 3 (neighborhood 2-3), balls centered at 5 and 6 (neighborhood 5-6), and balls centered at 6 and 7 (neighborhood 6-7) configure three new neighborhoods, as illustrated in panel A at Fig. 3-5. Following a similar reasoning, the intersection between neighborhoods 5-6 and 6-7 is not empty. Therefore, the union of these two neighborhoods configures a new neighborhood, particularly the union of balls centered at 5, 6, and 7. Following this procedure, i.e., the intersection of neighborhoods for identifying commonalities in shape, followed by their union for constructing novel neighborhoods, it is possible to delimit regions in the space with a similar shape.

This strategy results in groups of samples with a similar shape. However, these groups depend highly on the selected radii for the balls. For instance, when zero radii are selected, each group corresponds to a single sample because all ball intersections are empty. In contrast, for very large radii, for example, infinite, there is a single group containing all samples because all balls intersect. Observe also how groups (and neighborhoods) change when using a constant radius of 0.188 (panel A at Fig. 3-5) compared to a radius of 0.334 (panel C at Fig. 3-5).

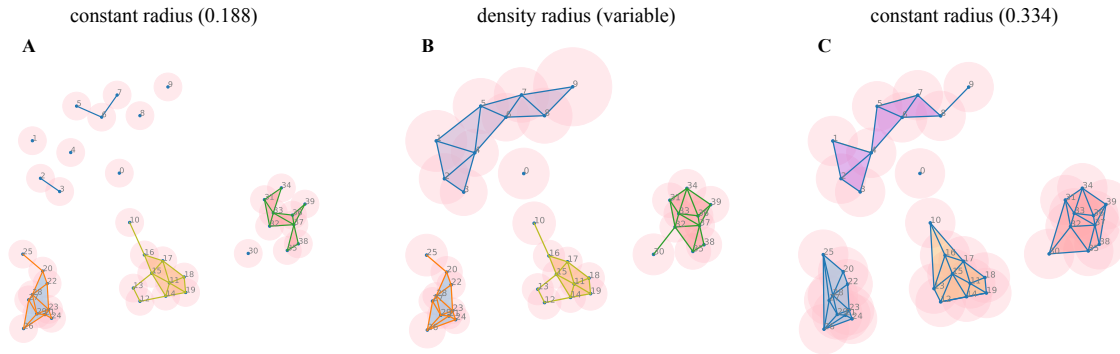


Figure 3-5: Comparison of balls of a constant radius with balls of radius dependent on the density. The figure presents three graphs composed of components connected by edges and triangles that rely on the radius of the balls. Connected components belonging to the same group are displayed in the same color. In panel A, the radius is small and constant for all samples. Note how the samples at the top of the graph have a lower density. These samples may correspond to the same category. However, with this radii selection they do not belong to the same neighborhood. In panel C the radius is larger and constant for all samples. In this case, the radii selection results in a neighborhood (violet) previously not observed in panel A. However, notice how a small increase in the radius will result in the mixture of orange and violet groups forming a single group, hiding the existence of the violet neighborhood. When a constant radius is used, the underlying assumption is that the samples have the same density in all groups. However, if some neighborhoods have samples with different densities, this assumption does not hold. In this case, using balls adapted to the local density can improve the delimitation of the neighborhoods. The method considers a large radius in regions of low local sample density, and a small radius in regions of high density, to discover the clusters, as observed in Panel B. In this case, the groups are more cohesive and robust to interclass samples.

Representing neighborhoods with simplicial complexes

In order to simplify computations, it is possible to characterize the neighborhood using a discrete structure called a *simplicial complex*, which provides a combinational representation of neighborhoods. A simplicial complex is a family of sets closed under subset operations. Each set in the simplicial complex will correspond to a part of the neighborhood.

For constructing the simplicial complex, first, each ball is identified by its center, i.e., by the index of the sample located in the ball center. These indices are included in the simplicial complex as singletons, i.e., a set with exactly one element. Each singleton is called a 0-simplex. The union of two connected balls (with non-empty intersections) in the neighborhood is indicated by a set containing the pair of indices of the two corresponding balls. Each

set of two elements is an edge (connecting two samples), also called a 1-simplex. For instance, neighborhoods 5-6 and 6-7 are identified by the 1-simplices $\{5, 6\}$ and $\{6, 7\}$, respectively. Importantly, these two neighborhoods share a non-empty intersection (the ball 6) that also belongs to the simplicial complex, as the 0-simplex $\{6\}$. The complete set of 1-simplices is also included in the simplicial complex. Following a similar idea, a set containing the corresponding three-ball indices represents the union of three balls with a non-empty intersection. Each set of three elements is a triangle (connecting with edges to three samples), also called a 2-simplex. The 2-simplices are also included in the simplicial complex. Note that each triangle contains the edges in the borders and samples in the corners, and by construction, they are also part of the simplicial complex. For instance, the union of balls 15, 16, and 17 is represented by the 2-simplex $\{15, 16, 17\}$. A simplicial complex containing this simplex should also include the 1-simplices $\{15, 16\}$, $\{16, 17\}$ and $\{15, 17\}$ and the 0-simplex $\{15\}$, $\{16\}$, and $\{17\}$.

Therefore, when considering any subset of a simplicial complex, the resulting set is also in the simplicial complex, i.e., the simplicial complex is closed under the subset operation. A similar idea allows the representation of high-order intersections or interactions using high-order simplices, such as tetrahedrons and their generalizations, the k -dimensional simplices [24]. Fig. 3-6 illustrates geometrically, with points, edges, and triangles, three different simplicial complexes obtained when considering different sample ball radii. In these cases, the simplicial complexes contain not only the triangles but also the edges conforming the triangles and the corner points, conforming these edges.

Simplex dimension and cohesion

The dimension of simplices in the simplicial complex provides information about the level of interaction or *cohesion* between constitutive samples. For instance, if a simplicial complex contains three 1-simplices, for example, $S_1 = \{\emptyset, \{A\}, \{B\}, \{C\}, \{A, B\}, \{B, C\}, \{C, A\}\}$, the amount of interaction of samples is lower than for $S_2 = \{\emptyset, \{A\}, \{B\}, \{C\}, \{A, B\}, \{B, C\}, \{C, A\}, \{A, B, C\}\}$. Because S_2 contains a higher order intersection between samples $\{A, B, C\}$, i.e., these three samples have more in common, exhibiting a stronger similarity relationship. In other words, samples in S_2 have more interactions among them than in S_1 . Similarly, if $S_3 = \{\emptyset, \{A\}, \{B\}, \{C\}, \{D\}, \{A, D\}, \{A, B\}, \{B, C\}, \{A, C\}, \{A, B, C\}\}$, then samples in S_3 have also lower cohesion than samples in S_2 because this simplicial complex includes an additional low dimension interaction $\{A, D\}$.

Importantly, evidence from experimental psychology suggests that the number of matching features between samples influences the degree of similarity in a category [45, 111]. Therefore, for defining categories, the level of cohesion in the simplices contained in the corresponding simplicial complex that represents the neighborhoods will be considered an indicator of this level of matching.

Categories and robustness

The category discovery method aims to characterize highly robust groups. However, groups of similar samples highly depend on selected radii. The proposed approach assumes that shape categories correspond to robust groups to overcome this limitation. The method’s definition of robustness draws on the TDA concept of persistence [24], which measures the level of constancy of data’s topological features across various scales. Specifically, our strategy seeks the most consistent number of groups, observable when the radii vary, i.e., the persistence in the number of groups. In addition, the method also aims for the groups with a minimal number of constituting samples q . This last requirement accounts for discovering groups with a representative number of samples. This parameter is problem dependent and can be selected by experts.

For characterizing groups with high persistence, first, let’s assume that the radii of the balls centered at each sample in the morphospace can be parametrized by a single value α , called the *filtration value*. The notion of filtration value aims to capture the scale of the neighborhoods. For this, the parametrization will be selected such that given two filtrations value α_1, α_2 with $\alpha_1 < \alpha_2$, the ball parametrized by α_1 will be contained in the ball parametrized by α_2 , for each ball centered at each sample, i.e., the scale described α_1 will be finer than α_2 .

The parametrization can be obtained, for instance, by using a $\alpha > 0$ as the radii for all balls (for example, see panels A and B in Fig. 3-5). However, other parametrizations are possible, as illustrated by Fig. 3-6, which shows a parametrization based on a filtration value dependent on the sample spatial density. We recall that balls centered at each sample (red balls in Fig. 3-6) are related to neighborhoods (union of intersecting red balls in Fig. 3-6). It is worth observing that with this parametrization, the neighborhoods related to small filtration values will always be contained in the ones associated with larger filtration values, i.e., the neighborhoods related to fine scales will be contained in coarse scales.

With this parametrization, a particular filtration value will help to determine a specific number of neighborhoods emerging at a certain scale. For instance, small filtration values may link to many neighborhoods, as illustrated by the red balls emerging for $\alpha = 0.057$ in Panel A in Fig. 3-6. In other words, when considering finer scales, many neighborhoods will appear. However, these neighborhoods are not necessarily groups because they do not have the minimal number of samples, in this case $q = 4$. In contrast, medium filtration values like $\alpha = 0.152$ (Panel C in Fig. 3-6), $\alpha = 0.178$ (Panel D in Fig. 3-6), and $\alpha = 0.26$ (Panel E in Fig. 3-6) will result in fewer neighborhoods, but some of them will contain enough samples to be considered as groups. Finally, large filtration values will determine only one group, as observed for $\alpha = 0.332$ at Panel F in Fig. 3-6.

As observed in Panel B, the number of groups obtained from this analysis can be studied as a function of the filtration value. For instance, the number of groups conformed by k -dimensional simplices emerging when increasing the neighborhood scale size from fine to

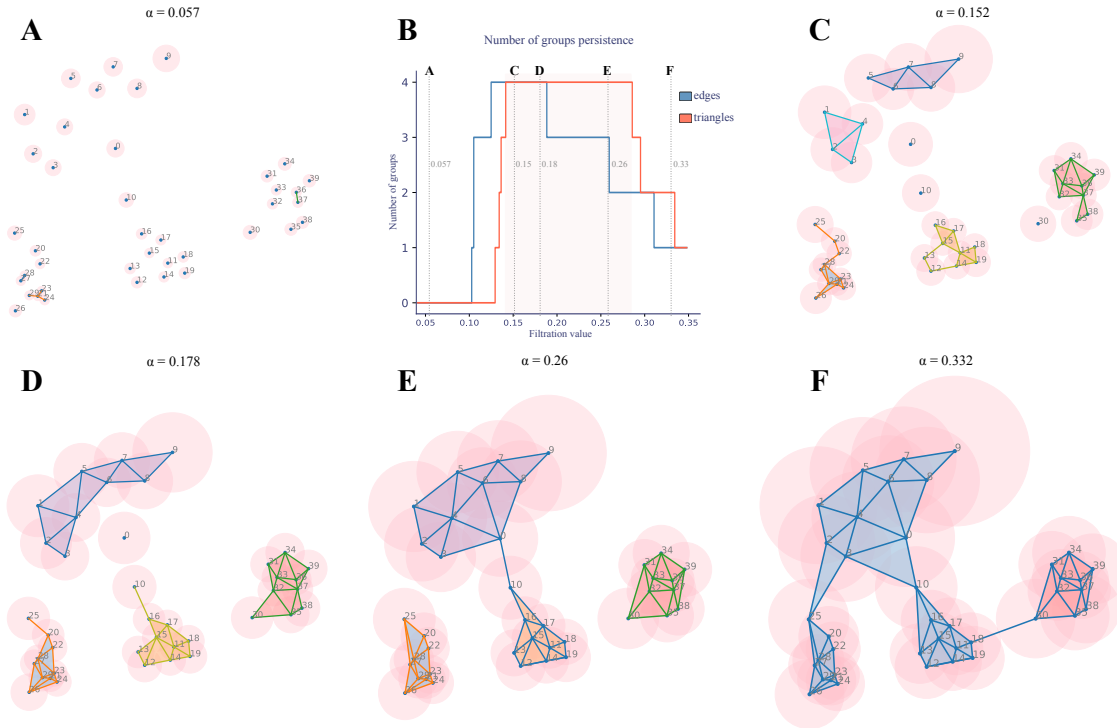


Figure 3-6: Groups emerging along different filtration values. The filtration value α determines the radii of spheres that are centered at each point in the morphospace. Similar shapes are captured in neighborhoods composed of the union of intersecting balls. These neighborhoods correspond to sets with points (0-simplices), edges (1-simplices), and triangles (2-simplices). Different filtration value result in neighborhoods with different scales. Small filtration values, for instance, Panel A, relate to fine-scale neighborhoods. Large filtration values, for instance, Panel F, relate to coarse-scale neighborhoods. When the filtration value increases, the scale of the neighborhoods also increases, and the number of groups emerging at different scales can be studied, as shown in Panel B. Five specific filtration values $\{A, C, D, E, F\}$ are indicated by the vertical dashed lines. In state A, many neighborhoods emerge, but only a handful of elements are interconnected, and no neighborhoods contain enough samples to be considered as groups. In panels C, D, and E, the same number of groups conformed by triangles emerge. In Panel F, only two groups formed by triangles emerged. The method searches the interval of filtration values showing the most consistent number of groups, where each group conformed by k -simplices of high-dimension, i.e., the largest persistent interval with the highest cohesion. The interval of filtration values with the greatest persistence is highlighted in light red.

coarse, as illustrated in Panel B of Fig. 3-6, which shows the number of groups conformed by lines (blue curve) or triangles (red curve). For characterizing robust groups, the method

focuses on the interval of filtration values that provides the most persistent number of groups, i.e., the largest interval of filtration values that always results in the same number of groups. When high dimensional simplices constitute these groups, they are both highly cohesive and robust across scales, then they are considered as categories. For instance, the light-red area in Panel B of Fig. 3-6 shows the interval of filtration values for which the groups constituted only by triangles showed the highest persistence number of groups. In this case, there are four categories.

Computing neighborhoods for categories

The previous approach computes cohesive and highly robust neighborhoods. However, for a set radius, the naive construction of the simplicial complex computing neighborhood by evaluating all combinations of possible intersections can be computationally intractable even for small samples. For instance, for 150 samples, there are more than half-million evaluations to be considered only to determine possible triangles in an exhaustive test of intersections of three balls.

For overcoming this problem, the proposed method precomputes the Delaunay triangulation (DT) of the samples and filters this simplicial complex, keeping only interactions (simplices) below a specific neighborhood scale, i.e., a particular filtration value. The DT is the simplicial complex corresponding to the geometric dual of the Voronoi diagram (VD) [82, 6]. In this case, the VD of the sample is a partition of the morphospace into regions close to each of a given set of samples. Once this partition is established, it is easy to query which samples are proximal in shape to a particular sample [36]. In particular, the ones connected by the Voronoi partition containing the sample [36, 82]. Because of this property and the duality, simplices in the DT relate proximal samples in shape, dramatically reducing the number of interactions to evaluate [47].

The DT, as a simplicial complex, represents a particular neighborhood that connects all samples. The aim is to remove some simplices of this simplicial complex to obtain cohesive neighborhoods. A cohesive neighborhood is a subset of the DT, which is also a simplicial complex, i.e., a sub-simplicial complex. For computing the sub-simplicial complexes corresponding to categories, the method should filter out some simplices of the DT. For instance, 1-simplices connecting two samples whose distance is bigger than the sum of the two ball radii should be discarded.

Additionally, the distance between two samples provides information about the neighborhood sizes for which two ball samples have an intersection, i.e., the exact size for which an edge or other high-order simplex may emerge. Therefore, to discard simplices in the DT is enough to focus on the distances between samples because the neighborhood will eventually change only for these distances. The advantage of this approach is that the number of distances to evaluate is finite. Because these distances may provide information to filter out complexes of the DT, they will be used as filtration values.

Fig. **3-6** shows the resulting neighborhood after discarding complexes from the DT using increasing filtering values. In this work, the Euclidean distance was used. Note how small filtration values filter most high-dimension simplicial complexes of the DT, keeping mainly 0-simplices (see Panel A at Fig. **3-6**). In contrast, large filtering values ($\alpha = 0.178$) result in sub-simplicial complexes that resemble categories, see Panel D at Fig. **3-6**). These sub-simplicial complexes are further post-processed to obtain highly cohesive neighborhoods with enough samples. For instance, four highly cohesive sub-simplicial complexes can be obtained by removing the 1-simplex $\{0, 10\}$ in Panel E at Fig. **3-6**.

Finally, the distances between points are ordered increasingly to construct the so-called persistence diagram (see Panel B at Fig. **3-6**), which shows the number of highly cohesive groups (for instance, edges or triangles) emerging for different filtration values. From this persistence diagram, the largest persistence interval is computed (see the light-red area Panel B at Fig. **3-6**), obtaining the robust categories.

Non-uniform density

In an ideal scenario, all samples in the same category appear at the same location in the morphospace. However, the biological sampling process and the morphological variability may cause a non-uniform distribution of samples. In other words, for real-world leaf traits, samples may have a non-uniform density [116]. Because categories depend on distances among points, when the filtering value increases, some samples in the same category may connect faster than others, as observed for samples 2 and 3 compared to sample 9 in Panel A at Fig. **3-5**, generating, for instance, incorrect categories. A factor that depends on the local density adjusts the distances between points to compensate for this phenomenon. More specifically, a kernel density estimation [124] with an exponential kernel function and bandwidth of $h = 0.5$ provides the local density of sample x_j as follows:

$$\hat{f}(x) = \delta_i = \frac{1}{Nh} \sum_{i=1}^N \exp\left(-\frac{(x_j - x_i)^2}{2h^2}\right) \quad (3-1)$$

with x_i the i -th sample.

Next, local density factors δ_i and δ_j computed for two samples x_i and x_j modify the distance $d_{i,j}$ between samples resulting in a transformed space, as follows:

$$d'_{i,j} = \frac{d_{i,j}}{f_i + f_j}$$

with $f_i = 1/e^{\delta_i}$ a compensation factor. This compensation factor always increases the size of the ball, but increases are larger for samples located in low-density regions. For instance, suppose that a point i is in a low region density. Then local density δ_i is small, and the corresponding compensating factor f_i is large. Therefore, all distances to i ($d'_{i,*}$) decrease

compared to the original distances ($d_{i,*}$), i.e., in the transformed space, the point i now is closer to all the other points, which is conceptually equivalent to enlarge the size of the ball around i in the original space.

Fig. **3-5** compares the filtering process using a constant radius (Panel A) and the proposed scaling factor (Panel B). Note that in all cases, the size of the balls increased in the transformed space (Panel B). However, increases were larger for samples located in low-density regions compared to the high-density areas, for instance, in sample 9 (low density) compared to 13 (high density).

Building taxonomies of visual categories

When the filtering value increases, different groups of samples or neighborhoods emerge. The proposed method looks for the most cohesive and robust ones for defining categories. Simplicial complexes representing neighborhoods at low filtering scales are always contained in simplicial complexes at high filtering scales (see Panels A, C, D, E, and F at Fig. **3-6**). This is a consequence of the filtering process performed on the DT. Therefore, following this nested sequence of neighborhoods, it is possible to track the construction process for the different categories.

A dendrogram showing how the groups are combined when the filtering value increases represents the category construction process from groups obtained by varying filtering values. This dendrogram allows a qualitative analysis of the proposed categories [65]. As a result, the selected shape trait displays which samples are most similar to each other as a taxonomy.

3.3.7 Experimental setting

The main objective of the evaluation is to compare the categories discovered by the proposed method with the categories established by experts. The proposed method operates unsupervised, receiving a sample and generating unlabeled categories. In order to assess its performance, a weighted multiclass f1-score is employed [113].

Initially, an algorithm matches the expert’s categories with the categories generated by the method. This match allows for a systematic comparison between the two sets of categories. Subsequently, a confusion matrix is constructed, where the samples that do not match the expert’s defined categories are assigned to X category. This approach ensures that all samples are accounted for and assessed in the evaluation. Precision and recall are calculated for each class to provide a more comprehensive evaluation. Precision reflects the accuracy of the method’s predictions within a specific category, while recall measures its ability to identify relevant instances within that category correctly.

The results are further weighted based on the size of each class to obtain the overall f1-score. This weighted approach considers the varying sizes of the categories, providing a balanced assessment of the method’s performance across all classes. Finally, the f1-score serves as a valuable metric to quantify the similarity between the expert’s category system

and the categories generated by the proposed method. This evaluation framework allows for a thorough and objective assessment of the method’s effectiveness in discovering categories without relying on expert labeling.

3.4 Results

This work proposes a novel robust shape category discovery approach for different leaf traits. First, we report evidence that the proposed strategy may automatically recover well-known categories of shapes for various leaf traits. Then, we compare the proposed method in the tasks of species discovery, assuming that each species corresponds to a shape category. This last task allowed evaluating leaf category discovery algorithms in previous works [116].

3.4.1 Discovering visual shape categories for leaf traits

Fig. 3-7 shows the dendrogram that summarizes the category discovery process in the apex shapes. In particular, this figure shows how different apex-shape groups and categories (vertical lines) emerged across different filtration values (vertical axis). The light-red area in the dendrogram indicates the interval with the highest persistence: 0.35-0.52, i.e., the one for which the number of categories was highly robust. For this interval, the proposed method automatically identified six shape categories. Significantly, these categories resemble the ones labeled manually, as observed in Fig. 3-8.

Dendrogram vertical lines in Fig. 3-7 indicate the emergence of different groups of shapes. At the bottom, the figure also shows shape samples corresponding to each group. Similar apex shapes constitute the categories when the filtration value increases. For instance, the first two contours from left to right (labeled as 0 and 43) have similar shapes. Therefore, the method join them, resulting in the first shape category. Similarly, the strategy combines the following four contours (labeled as 6 , 26 , 56 , and 55), resulting in the second shape category.

Fig. 3-9 shows the number of the groups identified versus the filtration value for edges (Panel A) and triangles (Panel B). This figure reports the persistence of groups formed by samples connected with edges and connected with triangles for each filtration value. It is worth recalling that the higher the dimension of the simplices conforming to the group, the higher the level of interaction among samples. Therefore, the group constituted by triangles is more cohesive than the group with edges. Fig. 3-9 highlights shaded areas with the most persistent groups for edges (light-blue area) and triangles (light-red area). As observed, in the case of groups conformed by edges, only two groups persist across filtration values, while for triangles, six groups emerged. ix categories were chosen in this case as this number provided the highest persistent value between edges and triangles, as observed in Panel C of Fig. 3-9.

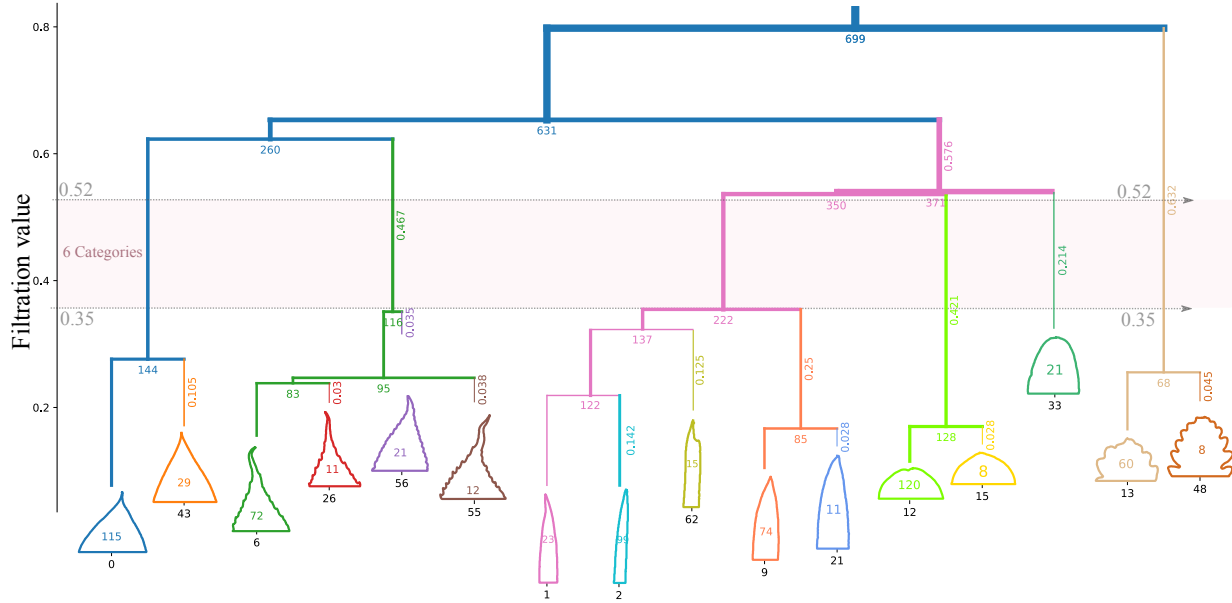


Figure 3-7: Dendrogram describing the shapes of leaf apices. This diagram organizes samples based on their respective leaf apex shapes. The horizontal axis shows the characteristic apex shapes within each category, complemented by the number of samples each shape represents. The vertical axis, on the other hand, reports the filtration value. The diagram demonstrates the organization of leaf apices into groups and the specific filtration value at which they are grouped. Each group's size, denoted by the line's width, indicates the number of samples within it. The value at the side corresponds to the persistence level. A light red region in the diagram indicates the groups with the highest persistence.

Fig. 3-8 reports the confusion matrix by comparing the resulting categories with the Hickey manual identified manually in the dataset. As observed, the proposed approach automatically recovered the original shape categories with an F1-score of 0.93. The method performed satisfactorily in most shape categories with a 5% error rate. Purple values in the matrix indicate these errors. In this case, *retuse* and *rounded* categories, which, as observed, are highly similar, resulted in the highest error rates. Column X details the number of samples per reference category; only 1.5% of samples were not classified. In addition, the method was highly robust to class imbalance. For instance, the *retuse* category has more than three times more samples than the *complex* category. However, the method performed appropriately in both.

Fig. 3-10 shows the morphospace formed by the first three principal components of the shape features used to describe the apex trait. A point in this space represents each leaf apex sample. The colors of the points indicate the six categories of the apex shape used as a reference, according to the Hickey manual. As observed from the sample distribution, categories can be recovered despite the differences in density and the varying number of













Categories							X	Total
 acuminate	113	0	8	0	0	0	3	124
 complex	0	68	0	1	0	0	0	69
 convex	0	0	131	0	0	0	1	132
 retuse	0	0	1	20	24	0	4	49
 rounded	0	0	0	0	103	0	1	104
 straight	0	0	0	0	1	218	2	221
Total	113	68	140	21	128	218	11	699

Figure 3-8: Confusion Matrix for the apex shape. This matrix compares the similarity between Hickey and the categories discovered by the algorithm. Rows show the apex shape categories from the Hickey manual [60]. Columns show the categories discovered by the proposed method. The contour of the most representative sample is shown as the column header. The column *Total* indicates the number of samples per category used in the experiment.

samples per category. In addition, this figure also illustrates the role of a proper shape description. For instance, the *complex* category resulted highly separated from the other shape categories (see PC2 vs. PC3 and PC1 vs. PC3), indicating that the proposed shape representation clearly distinguishes this category shape in the feature space. Furthermore, the figure shows how the category of *straight* can be eventually divided into two groups (see PC1 vs. PC2 and PC1 vs. PC3), potentially indicating the existence of subcategories within this class. Additionally, some samples of the *retuse* category are mixed with the ones in the *rounded* category, suggesting some overlap in the feature space between these classes and explaining the results in the confusion matrix.

Fig. 3-11 reports the margin (top panel) and the base leaf (bottom panel) traits' category discovery process. As before, the feature space and the dendrogram show how the categories emerged and illustrate some examples. At the same time, the confusion matrix compares the shape categories discovered by the proposed method against those used as a reference from the Hickey manual. The top panel in the figure summarizes the shape margin categories in the sample suggested by the algorithm. The corresponding distribution of points in the morphospace reveals that two groups may emerge in this sample. However, as observed, samples likely related to *entire* and *crenate* categories mixed, as did the samples of *dentate* and *serrate*. The dendrogram confirms this observation, showing how samples with similar margins conform to the groups. The dendrogram also shows how groups' margins are distributed from most serrated to most entire. The confusion matrix shows that these categories were mixed, which explains the F1-score of 0.65. Nevertheless, the matrix results

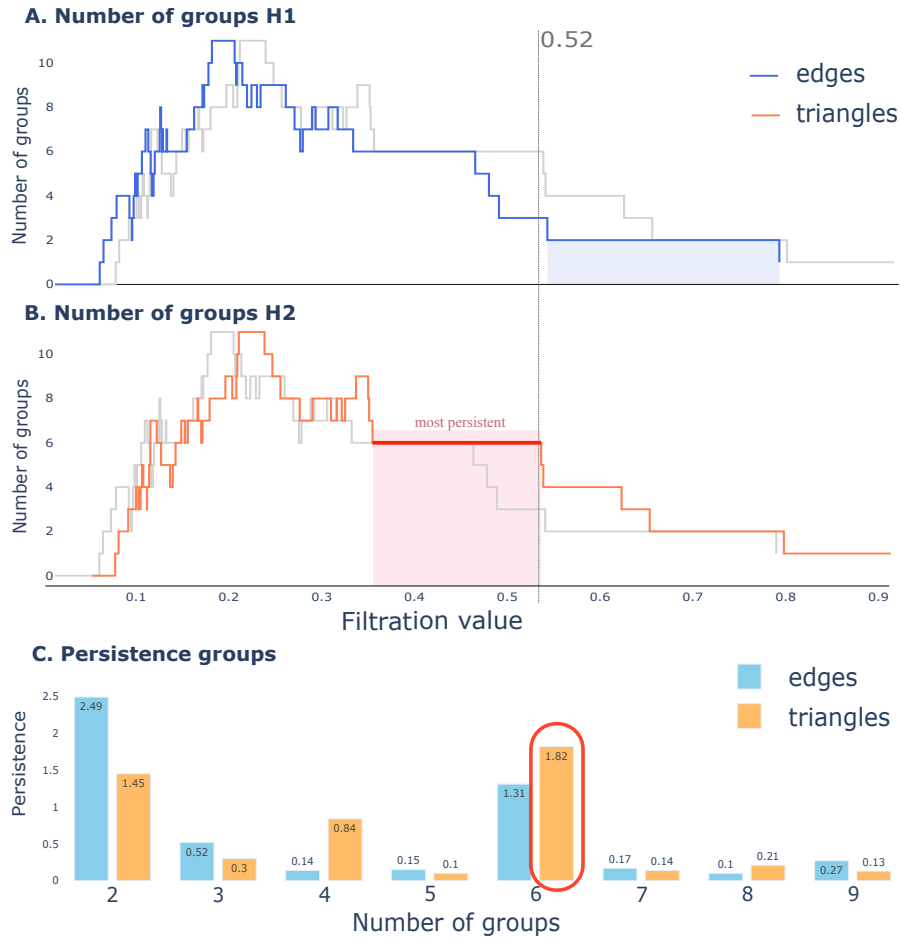


Figure 3-9: Number of identified groups across filtering values. Panel A shows the number of groups connected by edges for different filtering values. Panel B shows the corresponding number of groups connected by triangles. In both plots, the area representing the longest persistence is highlighted. Panel C displays the highest persistence value for both edge groups and triangles. In this case, 0.52 as the filtration value yields the most persistent value among triangle groups, resulting in six apex-shape categories.

are consistent with the morphospace distribution, indicating similarities in the feature space of these categories.

Fig. 3-11 at the bottom shows the shape category discovery results for the leaf base. The dendrogram shows how the groups were formed and mixed at different times. This dynamic nature of the groups' formation makes optimal category discrimination particularly challenging. In this case, the dendrogram revealed the existence of six distinct categories out of eight proposed by the Hickey classification scheme. Notably, two categories identified by the

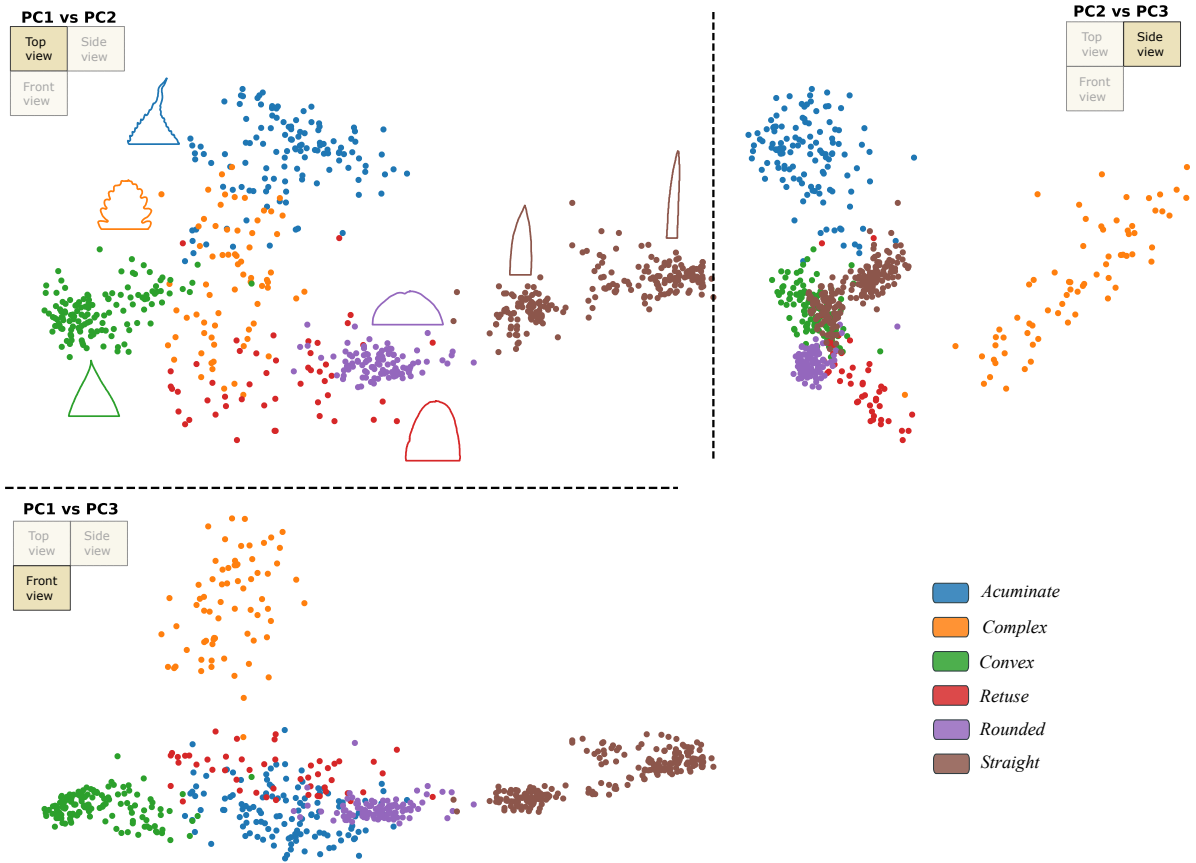


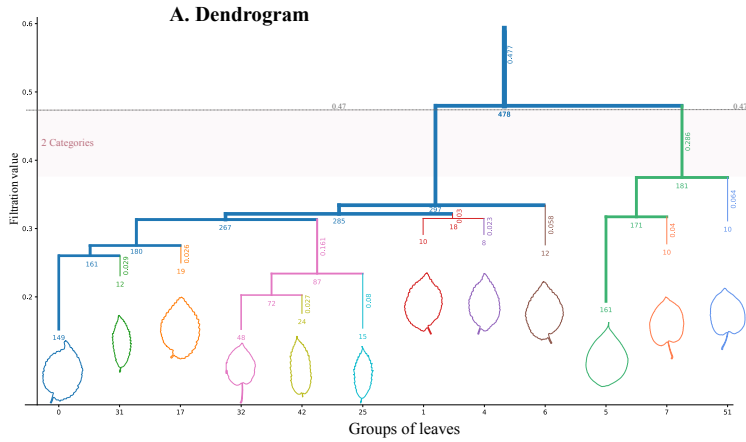
Figure 3-10: The morphospace generated from the features extracted from the leaf apex shape. Each point in the plot represents a selected sample leaf for this experiment. The colors of the points correspond to the manually annotated classes obtained using Hickey’s manual. A shape sample for each group is shown close to the discovered groups. The morphospace is generated based on the principal components (PC1, PC2, and PC3). Different groups of forms emerge in distinct regions within the morphospace.

dendrogram belong to the same class, as shown by their proximity in the dendrogram. Furthermore, the confusion matrix shows that three similar classes *cuneate*, *complex*, and *decurrent* merged, as also were *rounded* and *convex* classes. This decision likely was based on the observation that these shapes are highly similar, making it difficult to distinguish them. Significantly, the other categories were recovered satisfactorily.

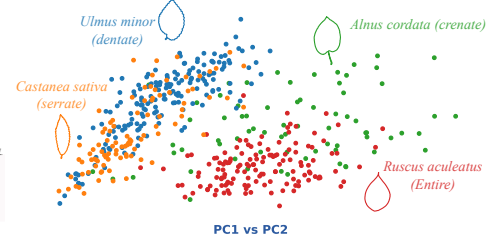
3.4.2 Discovering species as categories

Fig. 3-12 shows the results in the task of species discovery, i.e., unsupervised classification of leaves of the same species. This task implies discovering the proper categories and then

Margin type Analysis



B. Morphospace

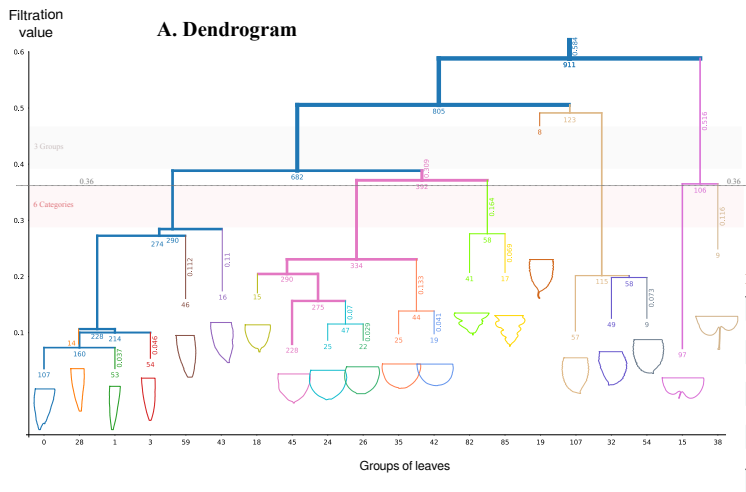


C. Confusion matrix

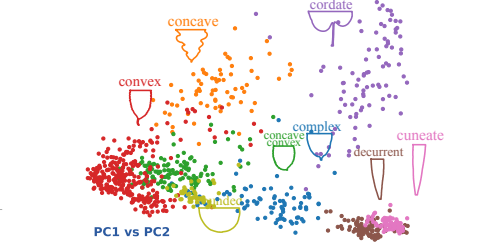
F1 score: 0.65

Margin-Type	Hickey	⊖	X	⊕	Total
	dentate	167	6	0	173
	serrate	94	2	0	96
	crenate	9	15	49	73
	entire	3	3	130	136
Total		273	26	179	478

Base-shape analysis



B. Morphospace



C. Confusion matrix

F1 score: 0.69

Base-shape Categories	⊖	⊕	X	⊗	Total					
	complex	7	0	0	0	68	0	9	0	84
	concave	0	53	0	0	0	0	16	0	69
	concave-convex	0	0	108	0	0	0	0	16	124
	convex	0	3	0	279	0	0	0	19	301
	cordate	0	0	0	0	86	1	9	3	104
	cuneate	0	0	0	0	0	98	0	0	98
	decurrent	0	0	0	0	0	0	84	0	84
	rounded	0	0	0	41	0	4	0	2	47
Total		7	56	108	320	86	255	9	65	911

Figure 3-11: Category discovery for margin and base. The top panel shows the categories discovered for the margin. The bottom panel shows the categories discovered for the base. Dendrograms (panels at A) represent the category discovery process across different filtration values. The light red area identifies the categories identified as the most persistent groups. Panels at B show the distribution of samples in the morphospace, where the colors indicate the ground truth category along with a representative sample. Panels at C show the results of the confusion matrix and the F1-score. The original categories are in rows, and the identified categories are in columns.

categorizing each sample correctly. The figure shows the results for eight species in the ImgeCLEF2012 dataset at the top panel. As observed, initially, the dendrogram suggested the existence of five shape categories or species. However, the dendrogram also highlights

a light-blue area with seven groups. When comparing the results of discovered species with the reference species, the method can successfully recover all species, as seen in the dendrogram’s light-blue shaded area. The high persistence area has seven lines, one for each species, and the left line corresponds to the mixture of two similar species. The confusion matrix confirmed this result, with six columns associated with one species and column C with the two remaining species. Only 5 of 905 samples were labeled as another species, and less than 8% (Column X) needed to be labeled for a $F1 - score = 0.82$, which shows high performance in the species discovery task.

Fig. **3-13** shows at the bottom the species discovered for a second dataset with 20 samples per species distributed in six species. In this case, the dendrogram shows that all species are connected almost simultaneously and remain so for a long interval of the filtration value. The quantitative results show perfect performance in the species recovery task and high visual similarity across species.

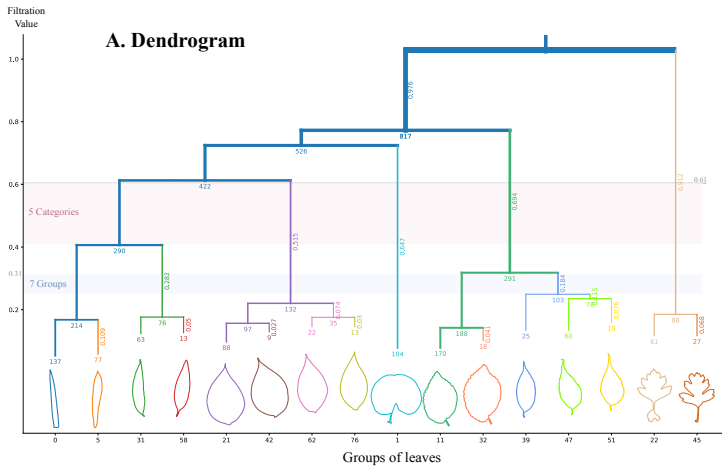
Finally, Fig. **3-13** summarizes all the shape categories discovered for the complete set of leaf traits studied. As observed, the proposed method correctly identifies most shape categories manually marked in the datasets for very different leaf traits and also for different species.

3.5 Discussion

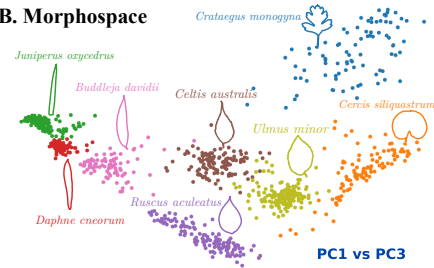
This paper proposes a novel approach for automatically discovering robust visual categories present in a sample of leaves. In contrast to previous works focused on leaf blades [116], the proposed strategy allows characterizing categories of shapes for other leaf traits, including apex, bases, and margins. The proposed approach successfully described visual categorical systems of shapes that highly resemble those independently identified by experts in classical botanic manuals, as observed in Fig. **3-13**. The strategy relies on two main components: first, a set of morphospaces suitable for representing shape features shared across similar leaf traits, and second, a TDA-based algorithm for identifying categories of shapes in these morphospaces. The computational strategy considers two properties humans use to define categories: 1) high levels of cohesion among samples, i.e., a high number of matching features (see Fig. **3-6**), which affect subjective similarity [45, 111], and 2) the existence of robust neighborhoods among samples belonging to the same category (see Fig. **3-6**), compared to related formal methods of categorization whose primary focus is similarity measurement [116, 101]. The strategy also accounts for the non-uniform distribution of samples in the morphospace, expected for real-world sampling processes [5, 105]. Additionally, the approach is highly interpretable [86], providing taxonomies for discovered visual categorical systems and interpretable representations of these categories in the morphospace. Results show that the proposed method may provide simultaneously qualitative (i.e., morphological features) and quantitative (i.e., categorical systems) descriptions of leaf morphology, representing a new alternative for both leaf shape description and knowledge discovery [127, 43, 120].

The discussion is organized into four parts. The first discusses the importance of defining

Species ImageCLEF Analysis



B. Morphospace

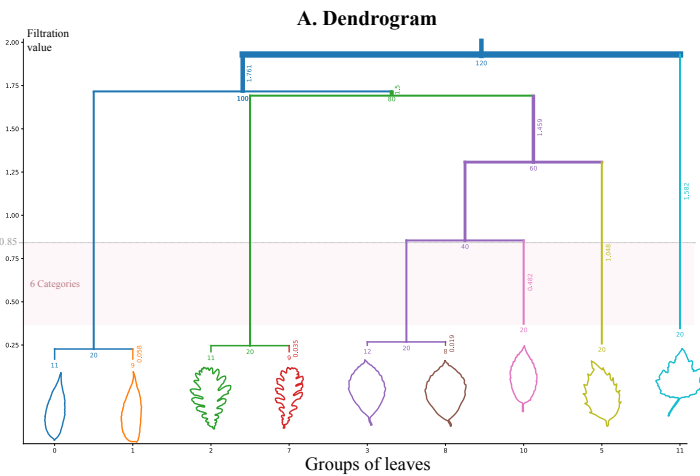


C. Confusion matrix

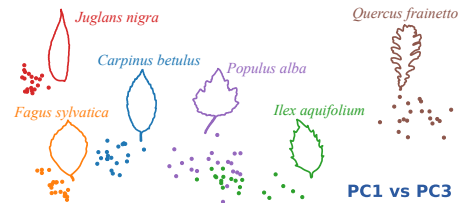
F1 score: 0.82

ImageCLEF Species				X				Total
<i>Crataegus monogyna</i>	71	0	0	17	0	0	0	88
<i>Cercis siliquastrum</i>	0	96	0	8	0	0	0	104
<i>Juniperus oxycedrus</i>	0	0	122	1	0	0	0	123
<i>Daphne cneorum</i>	0	0	81	3	0	0	0	84
<i>Ruscus aculeatus</i>	0	0	0	4	128	0	0	132
<i>Celtis australis</i>	0	0	0	18	0	102	0	124
<i>Buddleja davidii</i>	0	0	0	13	0	0	71	84
<i>Ulmus minor</i>	0	0	0	12	0	1	0	153
Total	71	96	203	85	128	103	71	905

TreeMEW Analysis



B. Morphospace



C. Confusion matrix

F1 score: 1.0

TreeMew Species							Total
<i>Carpinus betulus</i>	20	0	0	0	0	0	20
<i>Fagus sylvatica</i>	0	20	0	0	0	0	20
<i>Ilex aquifolium</i>	0	0	20	0	0	0	20
<i>Juglans nigra</i>	0	0	0	20	0	0	20
<i>Populus alba</i>	0	0	0	0	20	0	20
<i>Quercus freinetto</i>	0	0	0	0	0	20	20
Total	20	20	20	20	20	20	120

Figure 3-12: Category discovery for different species. The top panel shows the categories discovered for the ImageCLEF dataset. The bottom panel shows the categories discovered for the TreeMew. Dendrograms (panels at A) represent the category discovery process across different filtration values. The light red area identifies the categories identified as the most persistent groups. Panels at B show the distribution of samples in the morphospace, where the colors indicate the ground truth category along with a representative sample. Panels at C show the results of the confusion matrix and the F1-score. The original categories are in rows, and the identified categories are in columns.

which specific shape traits will be analyzed as a fundamental part of the category discovery methodology. The second section establishes the definition of category and group concerning state of the art and describes the method of category discovery and its properties. The third

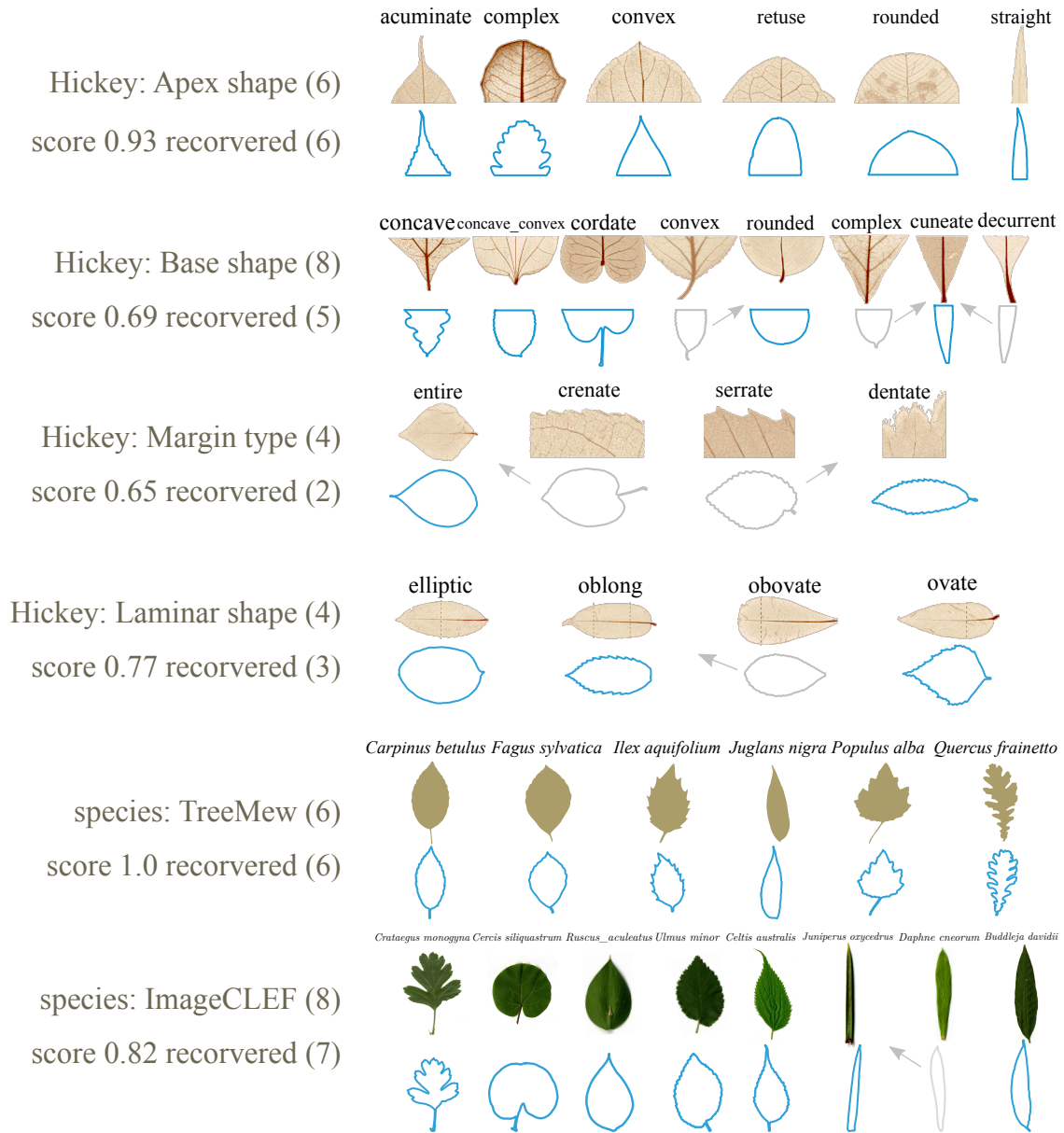


Figure 3-13: Shape categories identified for different traits and species. Results for six shape category recovery experiments are shown. The number of categories initially identified, the number of categories discovered, and the F1-score are reported. A reference image and the corresponding name are shown for each category to be discovered in each experiment. A contour representation of the corresponding recovered category is shown. The successfully recovered categories are highlighted with a blue contour. In contrast, missing categories are depicted as gray contours with a arrow indicating the category to which they were merged.

section describes the elements that allow interpreting the results in biological terms. Finally, limitations and future work are presented.

3.5.1 Importance of finding specific trait categories

Visual categorical systems in botanical manuals represent the gold standard tool for describing leaf shape and their traits. Plant experts widely utilize these descriptive systems to characterize leaf morphological variability, conduct taxonomic classifications, and even discover new species, among other biological description tasks [62, 80, 10, 120, 88]. These systems are essential in agriculture and ecology as they facilitate plant species identification. This task is crucial in agriculture for selecting and identifying the species that best adapt to environmental conditions [26, 100, 106]. In ecology, these systems play a vital role in understanding biodiversity, plant-environment interactions, and the impacts of climate change. Such understanding is instrumental in effectively conserving and managing ecosystems [3]. For instance, in a study for determining a new species, the expert consults visual categorical systems on botanical manuals, books, online databases, and herbariums to communicate and contrast a hypothesis. This work naturally has a high qualitative charge, such as qualitatively categorizing particular leaf traits, requiring a high degree of training [9, 107]. However, the expert should still account for quantitative description, aiming for objective measurements that provide evidence supporting the underlying hypotheses [108]. Therefore, the shape description process should be informative enough to enable communication and sufficiently accurate to provide convincing evidence. However, current approaches to shape description lack this harmonized view of qualitative and quantitative characterizations. The proposed method represents the first step in this direction. In particular, the results show how starting from non-annotated samples, it is possible to construct highly qualitative visual shape categorical systems (see Fig. 3-7) that can also be quantitatively described (see Fig. 3-8), conciliating both perspectives.

In addition, these qualitative and quantitative descriptions become increasingly more complex with the expanding amount of leaf information resulting from automatization [61], the increasing complexity of research questions [83], and the growing need for specialists [107]. Therefore, it is almost impossible to construct particular categorical systems to describe biological information for all different instance problems. Nevertheless, the results show that with minimal prior knowledge and no annotated information in the sample, the proposed system may automatically recover visual categorical systems similar to the ones offered by experts. In particular, the results (figures 3-8, 3-12, 3-11, and 3-13) demonstrated that it is possible to discover reasonably well the specific traits categories observed in the widely used Hickey manual and the original species in datasets samples. Notably, the same discovery strategy worked across six different leaf traits: base shape, apex shape, laminar shape, and margin type, and two datasets with various species (TreeMew and ImageCLEF) (see Fig. 3-13). These results confirm that it is possible to use computational approaches

to automatize the task of proposing visual categorical systems across highly heterogeneous biological settings, at least in the constrained domain of leaf morphology.

A critical aspect of a successful visual category discovery is the input-sample representation, as illustrated in Fig. **3-11**. Any categorical model assumes these inputs have some formal description [67], which define the morphospace [21]. These representations should capture the shape commonalities configuring a category while distinguishing among categories. Therefore, this work proposes particular representations accounting for shape differences in leaf traits. As expected, the method's effectiveness and degree of interpretability depend on the quality of these representations. For example, the proposed approach showed difficulties separating different margin types (Fig. **3-11**), presumably because of the limited capacity of the proposed representation to distinguish shape differences between margin categories. In this case, the proposed representation could be refined [21] to obtain categorical systems closer to the expert's expectations. After determining a more suitable description, the category discovery process can, for instance, be applied to other samples, decoupling the shape description tasks from the identification of categories. This property could help to identify potential new species.

3.5.2 Importance of defining robust categories

This work proposes an approach to generate robust categories from points in a representation space (Fig **3-10**). The notion of categories refers to a set of similar entities that are grouped together [49]. These groups simplify our worldview, providing a high explanatory capacity to the interpretations based on them [8]. For these reasons, the hierarchical organization of existing objects furnished by categories represents a fundamental mechanism of knowledge-building in biology, medicine, and other areas focused on complex phenomena description. The proposed approach distinguishes between group and category. Like clustering approaches, this approach focuses on similarity to form groups. However, to define a category, starting from groups based on similarity, three well-known concepts from topology and computational topology are considered [24]: neighborhood, cohesion, and persistence. The neighborhoods allow us to describe the extent of the closest points to each sample in the morphospace, providing a well-grounded approach for defining similarity. This concept enables us to characterize, for instance, more abstract (large filtration values) to more concrete (small filtration values) groups, see dendrogram in Fig. **3-7**, similar to the groups proposed in the biological taxonomies [40]. Cohesion permits considering groups with many commonalities among features, as observed in panels A and B in Fig. **3-6**. This characteristic influences the subjective perception of categories [45, 111]. Persistence measures how neighborhood relationships among samples endure or survive across multiple scales or sizes of neighborhoods [24]. This last concept provides a grounded alternative to define categories by looking for the most persistent groups, see the light-red area in Fig. **3-7**. Remarkably, in most cases, the resulting groups characterized using these three topological properties coincide with the

categories provided by experts (for instance, see the confusion matrix in Fig. 3-8). Most recent works heavily focused on learning representations spaces using data-intensive strategies, such as deep learning [23] or sensitivity to minor sample shape variations and outliers [116]. Nevertheless, to our knowledge, this work is the first to consider these three properties to characterize groups as categories. The evidence highlights the importance of considering additional properties that likely describe the category concept beyond the ambiguous notion of similarity.

The proposed method is highly robust to non-uniform densities in the sample groups of the morphospace. This non-uniformity may result from extracting feature samples with some variability, as illustrated in Fig. 3-5. This variability may result from the sampling process, for instance, when sampling leaves at different stages of development [42] or if samples are affected by insects [18]. Estimating groups under non-uniform density conditions is a well-established problem commonly faced in density-based clustering methods [89]. Nevertheless, in the proposed approach, this non-uniformity may also affect the persistence of groups and, therefore, the category discovery process [116]. The method considers a factor that depends on the statistical density at each point and modifies the local neighborhood's size to compensate for the non-uniform densities in the morphospace. This factor enlarges the neighborhood in high-density areas and decreases it in low-densities regions (Fig. 3-6). This strategy allows the method to operate with any number of samples, overcoming the limitation of class balance required by some clustering methods [75], as long as the groups have enough samples. Introducing this compensation in the neighborhood size is also crucial because the strategy would fail without information on how many samples each group had.

3.5.3 Results interpretability

It is worth recalling that the category concept underlying the proposed method goes beyond the notion of group or cluster since it also accounts for cohesion and robustness based on varying neighborhoods. Interestingly, these neighborhoods and the underlying representations also allow the method to reach a high biological interpretability level [86, 81]. For instance, the neighborhoods configuring a category are easily interpretable in the proposed approach, as illustrated in Fig. 3-7. This interpretability based on neighborhoods contrasts with the lack of transparency of recent clustering approaches that learn non-linear representations from the samples to discover categories [23, 78, 121, 63], which are hard to interpret by experts. In addition, the cluster concept can be intrinsically ambiguous because different criteria for clustering data may result in different groups. For example, in principle, it is not known which trait defines the categorization's nature, and depending on the similarity criteria's extent, multiple groups can be obtained [102, 23]. The proposed approach still relies on representation selection to construct the morphospace. However, compared with previous proposals, our approach generates robust categories optimal across different scales, compared to iterative methods that require the optimization of objective functions, which

are not related to the biological nature of the problem [23, 38]. This is the case, for example, of methods based on neural networks [130], in which the reasons for observing a particular category may be obscure.

The proposed method provides highly interpretable groups and categories since the samples' relationship depends on their similarity and cohesion (Fig. 3-6). In the proposed construction, two samples are directly connected if there is a connecting path in the underlying simplicial complex. These paths also have an associated level of cohesion since they can be edges, triangles, or higher-dimension structures with a higher dimension. Samples with a high cohesion among them are characterized by identifying the dimension of the combinatorial structure. Importantly, to overcome the computational problem of comparing multiple subsets of samples, the proposed approach restricts the exploration of possible interactions to the connections part of the DT [47]. This restriction guarantees that the connection between points performs through the contiguous neighborhoods. Furthermore, this triangulation also decreases the method's computational complexity because a point is only connected to its immediate neighbors in this simplicial complex. This coherence in the similarity relationship also helps improve the results' interpretability.

The proposed approach provides high levels of interpretability from the results observed in the morphospace and the dendrogram. For example, the morphospace projected to 3D (Fig. 3-10) shows the samples' distribution density and variability. This figure also shows how the sample is organized for a specific neighborhood size, as seen in Figure 3-10. This visualization allows us to understand how the groups are conformed, how cohesive they are, and where the categories emerge. Additionally, the dendrogram (Fig. 3-7, 3-12, 3-11) quantitatively and qualitatively shows the whole dynamics of the method through changes in the size of the neighborhood. As observed, this approach explains how the groups appear, for instance, when they merged, with which sample groups they merged, how many samples each group accumulated, and the stability of each group. Remarkably, it is also possible to discriminate which groups, subgroups, and categories emerged. All this information is critical for knowledge communication.

3.5.4 Limitations and future work

The proposed approach can be extended in several ways. First, to increase the number of categories that can be identified, it's essential to consider new feature extraction algorithms. While this study examined six categories, future work may explore complementary representations. This exploration requires expanding the database to cover the requisite shapes sufficiently with enough specimens. Second, subsequent studies may explore possible correlations between the categories discovered with biological explanatory variables, for instance, investigating which morphological traits correlate best with these variables.

4 Conclusions and recommendations

4.1 Contributions

The main contribution of this thesis is the proposal of a computational model to discover objective, robust, and interpretable visual categories automatically. Unlike other automatic approaches, this model does not require annotations in the sample and can find visual categories with minimal prior knowledge. This approach marks a beginning toward bringing automated morphological descriptions closer to biologists, making them adaptable in their workflow. The proposed model synthesizes results or knowledge into robust categories based on the representation of the target trait. Therefore, biologists can formulate and test their working hypotheses using a similar strategy established in geometric morphometrics [16].

The proposed model produces quantitative results, such as categories with their cohesion and similarity values, as well as qualitative explanations, like data distribution in the morphospace and the formation of groups and categories in the dendrogram. By merging both perspectives, the model provides a comprehensive and insightful analysis.

By automating the categorization process, the model frees biologists from the laborious and time-consuming task of manual annotations. This automation enables them to focus on higher-level interpretations and hypothesis testing based on the generated visual categories. The model's flexibility allows it to adapt to different morphological features and research contexts. Biologists can readily incorporate their expertise into the model, tailoring it to their needs and enhancing its interpretability.

Furthermore, the model's ability to provide quantitative and qualitative outputs bridges the gap between objective data analysis and meaningful biological interpretations. This dual perspective enriches the biologist's understanding of shape variation and facilitates the integration of the automated approach into their current morphological studies.

The morphospace displays the distribution of samples (ideally using a reduced set of characteristics) and their relative variability. At the same time, the dendrogram provides taxonomies tailored to the samples by hierarchically organizing them. Within the morphospace, it is possible to introduce different variables for analysis, visualizing convex hulls alongside the categories and observing their correlations [116]. This versatility allows for exploring

relationships with external factors such as climate, ecology, soil characteristics, environmental pressures, and crop production [10]. These connections can explain and support global warming, climate change, environmental adaptations, functional ecology, morphological variability, and more [90].

An important aspect is handling unclassified specimens, denoted as category 'X' in the confusion matrices [96]. These unique forms within the sample can lead to significant ecological or biodiversity findings, such as discovering new species, identifying anomalies (diseases, herbivory), and characterizing biodiversity [77]

This approach holds immense potential for advancing our understanding of complex ecological processes and their connections to shape variations [122]. For instance, investigating how climate factors influence morphological diversity or how environmental pressures impact species adaptations can lead to valuable discoveries and contribute to broader environmental and ecological studies [26].

It is essential to highlight that the method identifies the optimum as the highest persistence at the specified level of cohesion. Unlike specific approaches such as deep clustering or novel category discovery, our method avoids using iterative methods to optimize objective functions that may not be directly relevant to the morphological description problem [63, 78].

Various approaches utilize the concept of category or propose the discovery of categories, although a formal definition of this concept often needs to be provided. While researchers commonly work with an intuitive understanding of what constitutes a category, there is, to our knowledge, no formal definition that allows for systematic development based on this concept.

In this study, we clearly define the visual category rooted in projecting a sample of elements in a representation space: *“a set of groups that persist at different scales and exhibit high cohesion.”* This definition is supported by similarity neighborhood, element cohesion, cohesion persistence, and non-uniform density in both space and groups, thus rendering the category concept more robust than a cluster. Additionally, incorporating cohesion persistence ensures that the identified categories are meaningful but robust and reliable across different scales.

On the other hand, the representation and category discovery components are independent. The first component encompasses various algorithms that extract crucial features in the analysis and generate a morphospace. In contrast, the second component solely relies on a given representation space. This characteristic allows the method to be applied to multiple problems requiring category discovery by simply changing the first component. For instance, in different contexts, it is commonly utilized for various purposes such as customer

segmentation to enhance personalized services and marketing campaigns, social network analysis to identify communities and behavioral patterns, anomaly detection for spotting fraudulent activities, text clustering for topic analysis or content recommendation, and notably, in medical image classification where interpretability is of utmost importance [125, 96].

Importantly, the main findings resulting from this work were also relevant for other categorization tasks explored during this project. For instance, the study titled “Highly Seasonal Aggressive Behaviors Link to Temporal Dynamics Shared Across Space” [117] aimed to identify categories of locations based on crime levels with a specific focus on seasonality. This approach uses a representation of crime data based on spectrograms of different zones within the city. The research explored the spatiotemporal patterns of aggressive behaviors. Subsequently, distinct categories of city zones exhibiting similar crime behaviors were established. In addition, in work titled “Spatial-temporal patterns of aggressive behaviors: A case study in Bogotá, Colombia” [118] constructed a comprehensive representation of crime by generating spectrograms based on the city’s zones. These spectrograms effectively characterized the spatiotemporal occurrences of criminal activities across the urban landscape. This representation helped to identify and group city zones exhibiting similar patterns of crime behavior into distinct categories.

In addition to the visual category discovery model, several original computational methods are provided that can be applied to various tasks:

- An algorithm for leaf image binarization, enabling precise segmentation of leaves in images captured from the front and under controlled backgrounds.
- A method for segmenting the petiole and leaf parts from a binary image, allowing for more detailed analysis and characterization.
- A technique to determine a local density adjustment factor for scaling the neighborhood size, facilitating the customization of analysis based on specific spatial requirements.
- Five distinct algorithms for extracting specific features from shape traits are employed for comparison with the categories specified in the Hickey manual.
- An optimization method for connecting components of multiple dimensions through Delaunay triangulation, enhancing data visualization and analysis.
- These innovative computational methods extend beyond the scope of visual category discovery and offer valuable tools for tackling various challenges in botany and shape analysis.

As a future direction, the inclusion of more algorithms for extracting specific features and validation using appropriate representation space metrics should be considered. An alternative to fixed reference systems like botanical manuals is establishing a standardized library

of algorithms recognized and endorsed by the biologist community for extracting and representing morphological traits. While there is still much work to be done in automating morphological description, this research is an initial step in bridging the gap between quantitative and qualitative aspects, allowing experts to engage in analysis and interpretation. Nevertheless, collaboration with expert biologists is essential for this approach's full potential.

The validation and refinement of this work should be pursued through close cooperation with specialists who possess domain knowledge. Expanding the repertoire of feature extraction algorithms, the proposed model can cover a broader range of morphological traits and address more diverse research questions. Validating the representation space through appropriate metrics will further enhance the model's reliability and applicability, strengthening its potential as a powerful tool for morphological analysis.

4.2 Conclusions

This work accomplished the objectives of developing a robust computational model for automatically discovering leaf shape categories while ensuring interpretability. The proposed representation space and category discovery strategy allow for meaningful and reliable results, paving the way for new insights and discoveries in leaf morphology. The proposed method successfully describes categorical systems that closely resemble those independently identified by experts in classic botanical manuals. Furthermore, including a transparent and interpretable explanation strategy enhances the model's applicability and fosters collaboration between computational researchers and domain experts, fostering advances in automated morphological analysis for various biological studies.

The foremost accomplishment was developing a computational model to discover leaf shape categories automatically. This model employs state-of-the-art algorithms and techniques to ensure accurate category discovery, proposing a suitable representation space tailored to the specific characteristics of leaf traits. This representation space involved defining essential morphological features that capture the diversity and variations present in the leaf samples. By selecting a contractive representation space that maximizes the discrimination power of these features, we improved the model's ability to distinguish between different leaf traits effectively.

One of the critical challenges in this research was to design a category discovery strategy that would be both robust and objective. To address this, we implemented a novel approach that does not rely on manual annotations in the sample data. Instead, our model autonomously identifies visual categories, significantly reducing the need for laborious manual efforts and

minimizing human biases. Researchers can gain insights into leaf shape's underlying patterns and structures by leveraging this approach. Identifying such categories offers a deeper understanding of the diverse morphological variations in the leaf sample, which is crucial for ecological studies, taxonomic research, and understanding the adaptability and biodiversity of plant species.

Transparency and interpretability are paramount when dealing with complex computational models in scientific research. Therefore, we meticulously designed a strategy to explain the shape categories discovered by our model in an interpretable way. The model produces quantitative results, such as cohesion and similarity values for each category, and qualitative explanations, including visual representations of the data distribution in the morphospace and the formation of groups in the dendrogram. This comprehensive analysis allows biologists and researchers to effectively understand and validate the categorization results. Unlike traditional methods that may produce abstract or complex categorizations, this approach generates visually interpretable categories, facilitating a more straightforward interpretation and communication of the results.

The method's versatility and ability to automatically discover meaningful visual categories make it a powerful tool for various applications beyond botanical research. Its potential applications extend to fields such as pattern recognition, computer vision, and artificial intelligence, where visual categorization plays a crucial role.

4.3 Limitations

The limitations of this work highlight specific areas that can be further explored and expanded upon in future research:

- **Limited number of analyzed traits:** The study is restricted to investigating only five leaf shape traits. Expanding the analysis to include a more extensive range of morphological characteristics can offer a more comprehensive understanding of leaf diversity and give researchers a broader scope for their investigations. Exploring additional traits can unveil hidden patterns and correlations that may contribute to a deeper understanding of plant taxonomy and adaptation.
- **Controlled sample background:** The model's current limitation to receiving leaf samples with controlled backgrounds may restrict its applicability to real-world scenarios. Future research could explore integrating images with natural backgrounds, which better emulate the complexities encountered in ecological settings.

- **Limited sample types:** The model's testing has been conducted solely using individual images of complete leaves, neglecting the analysis of leaflets. Including leaflets in the study can provide insights into the variations within compound leaves, contributing to a more comprehensive understanding of leaf morphology in diverse plant species.
- **Focus on shape traits:** While the current model successfully analyzes shape traits, it does not encompass venation patterns. Developing specialized algorithms to extract and analyze venation characteristics can overcome this limitation.

Bibliography

- [1] ADAMS, Dean C. ; ROHLF, F J. ; SLICE, Dennis E.: Geometric morphometrics: ten years of progress following the “revolution”. In: *Italian Journal of Zoology* 71 (2004), Nr. 1, S. 5–16
- [2] ALPAYDIN, Ethem: *Introduction to Machine Learning*. 2nd. Cambridge, Massachusetts, USA : The MIT Press, 2010
- [3] AMÉZQUITA, Erik J. ; QUIGLEY, Michelle Y. ; OPHELDERS, Tim ; MUNCH, Elizabeth ; CHITWOOD, Daniel H.: The shape of things to come: Topological data analysis and biology, from molecules to organisms. In: *Developmental Dynamics* 249 (2020), Nr. 7, S. 816–833
- [4] AMLEKAR, Manisha M. ; GAIKWAD, Ashok T.: Plant classification using image processing and neural network. In: *Data Management, Analytics and Innovation*. New York, NY : Springer, 2019, S. 375–384
- [5] ANDERSON, David R.: *Guidelines for line transect sampling of biological populations*. USA : The Unit, 1976 (9-76)
- [6] AURENHAMMER, Franz ; KLEIN, Rolf ; LEE, Der-Tsai: *Voronoi diagrams and Delaunay triangulations*. Singapore : World Scientific Publishing Company, 2013
- [7] BALL, H. ; EXELL, A. ; HARDING, J. ; L’EONARD, J ; LEWIS, J. ; MELDERIS, A. ; MELVILLE, R. ; STAFLEU, F. ; WALTERS, S. ; DUVIGNEAUD, P. ; PETIT, E. ; TOURNAY, R. ; DER VEKEN, P. V.: Systematics association committee for descriptive biological terminology. II. Terminology of simple symmetrical plane shapes (Chart 1). In: *Taxon* 41 (1962), Nr. 11, S. 145–156
- [8] BARITE, Mario G.: The notion of category: its implications in subject analysis and in the construction and evaluation of indexing languages. In: *KO KNOWLEDGE ORGANIZATION* 27 (2000), Nr. 1-2, S. 4–10
- [9] BATUT, Bérénice ; HILTEMANN, Saskia ; BAGNACANI, Andrea ; BAKER, Dannon ; BHARDWAJ, Vivek ; BLANK, Clemens ; BRETAUDEAU, Anthony ; BRILLET-GUÉGUEN, Loraine ; ČECH, Martin ; CHILTON, John [u. a.]: Community-driven data analysis training for biology. In: *Cell systems* 6 (2018), Nr. 6, S. 752–758

- [10] BAUMGARTNER, Aly ; DONAHO, Michaela ; CHITWOOD, Daniel H. ; PEPPE, Daniel J.: The influences of environmental change and development on leaf shape in *Vitis*. In: *American journal of botany* 107 (2020), Nr. 4, S. 676–688
- [11] BEENTJE, Henk: *The Kew Plant Glossary: An Illustrated Dictionary of Plant Terms*. Richmond, London, UK : Kew Publishing, Royal Botanical Gardens, Kew, 2010
- [12] BENDER, Amanda L. ; CHITWOOD, Daniel H. ; BRADLEY, Alexander S.: Heritability of the structures and ^{13}C fractionation in tomato leaf wax alkanes: a genetic model system to inform paleoenvironmental reconstructions. In: *Frontiers in Earth Science* 5 (2017), S. 47
- [13] BERDUGO-LATTKE, Mary L. ; GÓNZALEZ, Fabio ; RANGEL-CH, J O. ; GÓMEZ, Francisco: P-type based dimensionality reduction for open contours of Colombian Páramo plant species. In: *Ecological Informatics* 36 (2016), S. 1–7
- [14] BIOT, Eric ; CORTIZO, Millán ; BURGUET, Jasmine ; KISS, Annamaria ; OUGHO, Mohamed ; MAUGARNY-CALÈS, Aude ; GONÇALVES, Beatriz ; ADROHER, Bernard ; ANDREY, Philippe ; BOUDAUD, Arezki [u. a.]: Multiscale quantification of morphodynamics: MorphoLeaf software for 2D shape analysis. In: *Development* 143 (2016), Nr. 18, S. 3417–3428
- [15] BOGGESE, Albert ; NARCOWICH, Francis J.: *A first course in wavelets with Fourier analysis*. Hoboken, NJ, USA : John Wiley & Sons, 2015
- [16] BOOKSTEIN, Fred L.: *Morphometric tools for landmark data: geometry and biology*. Cambridge, UK : Cambridge University Press, 1997
- [17] BOR, Norman L.: *Manual of Indian forest botany*. Oxford, England, UK : Oxford University Press.; Geoffrey Cumberlege, 1953
- [18] BROWN, VK ; LAWTON, John H.: Herbivory and the evolution of leaf size and shape. In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 333 (1991), Nr. 1267, S. 265–272
- [19] BRYSON, Abigail E. ; WILSON BROWN, Maya ; MULLINS, Joey ; DONG, Wei ; BAHMANI, Keivan ; BORNOWSKI, Nolan ; CHIU, Christina ; ENGELGAU, Philip ; GETTINGS, Bethany ; GOMEZCANO, Fabio [u. a.]: Composite modeling of leaf shape along shoots discriminates *Vitis* species better than individual leaves. In: *Applications in plant sciences* 8 (2020), Nr. 12, S. e11404
- [20] BUCKSCH, Alexander ; ATTA-BOATENG, Acheampong ; AZIHO, Akomian F. ; BATTOGTOKH, Dorjsuren ; BAUMGARTNER, Aly ; BINDER, Brad M. ; BRAYBROOK, Siobhan A. ; CHANG, Cynthia ; CONEVA, Viktoirya ; DEWITT, Thomas J. [u. a.]: Mor-

- phological plant modeling: unleashing geometric and topological potential within the plant sciences. In: *Frontiers in plant science* 8 (2017), S. 900
- [21] BUDD, Graham E.: Morphospace. In: *Current Biology* 31 (2021), Nr. 19, S. R1181–R1185
- [22] CAL, Andrew J. ; SANCIANGCO, Millicent ; REBOLLEDO, Maria C. ; LUQUET, Delphine ; TORRES, Rolando O. ; MCNALLY, Kenneth L. ; HENRY, Amelia: Leaf morphology, rather than plant water status, underlies genetic variation of rice leaf rolling under drought. In: *Plant, cell and environment* 42 (2019), S. 1532–1544
- [23] CARON, Mathilde ; BOJANOWSKI, Piotr ; JOULIN, Armand ; DOUZE, Matthijs: Deep clustering for unsupervised learning of visual features. In: *Proceedings of the European conference on computer vision (ECCV)*, 2018, S. 132–149
- [24] CHAZAL, Frédéric ; MICHEL, Bertrand: An introduction to topological data analysis: fundamental and practical aspects for data scientists. In: *Frontiers in artificial intelligence* 4 (2021), S. 667963
- [25] CHIGNELL, Mark ; WANG, Lu ; ZARE, Atefeh ; LI, Jamy: The Evolution of HCI and Human Factors: Integrating Human and Artificial Intelligence. In: *ACM Transactions on Computer-Human Interaction* (2022)
- [26] CHITWOOD, Daniel H.: The shapes of wine and table grape leaves: An ampelometric study inspired by the methods of Pierre Galet. In: *Plants, People, Planet* 3 (2021), Nr. 2, S. 155–170
- [27] CHITWOOD, Daniel H. ; MULLINS, Joey: A predicted developmental and evolutionary morphospace for grapevine leaves. In: *Quantitative Plant Biology* 3 (2022), S. e22
- [28] CHITWOOD, Daniel H. ; OTONI, Wagner C.: Morphometric analysis of Passiflora leaves: the relationship between landmarks of the vasculature and elliptical Fourier descriptors of the blade. In: *GigaScience* 6 (2017), Nr. 1, S. 1–13
- [29] CHITWOOD, Daniel H. ; SINHA, Neelima R.: Evolutionary and environmental forces sculpting leaf development. In: *Current Biology* 26 (2016), Nr. 7, S. R297–R306
- [30] COUSSEMENT, Jonas ; STEPPE, Kathy ; LOOTENS, Peter ; ROLDÁN-RUIZ, Isabel ; DE SWAEF, Tom: A flexible geometric model for leaf shape descriptions with high accuracy. In: *Silva Fennica* 52 (2018), Nr. 2
- [31] DAELLI, Valentina ; VAN RIJSBERGEN, Nicola J. ; TREVES, Alessandro: How recent experience affects the perception of ambiguous objects. In: *Brain research* 1322 (2010), S. 81–91

- [32] DE MAESSCHALCK, Roy ; JOUAN-RIMBAUD, Delphine ; MASSART, Désiré L: The mahalanobis distance. In: *Chemometrics and intelligent laboratory systems* 50 (2000), Nr. 1, S. 1–18
- [33] DOSHI-VELEZ, Finale ; KIM, Been: Towards a rigorous science of interpretable machine learning. In: *arXiv preprint* (2017). – <https://arxiv.org/abs/1702.08608>
- [34] ELДАР, Yonina C. ; OPPENHEIM, Alan V.: MMSE whitening and subspace whitening. In: *Information Theory, IEEE Transactions on* 49 (2003), Nr. 7, S. 1846–1851
- [35] ELLIS, Beth ; DALY, Douglas C. ; HICKEY, Leo J. ; JOHNSON, Kirk R. ; MITCHELL, John D. ; WILF, Peter ; WING, Scott L.: *Manual of leaf architecture*. Ithaca, NY, USA : Cornell University Press Ithaca, 2009
- [36] ERWIG, Martin: The graph Voronoi diagram with applications. In: *Networks: An International Journal* 36 (2000), Nr. 3, S. 156–163
- [37] FAILMEZGER, Henrik ; LEMPE, Janne ; KHADEM, Nasim ; CARTOLANO, Maria ; TSIANTIS, Milto ; TRESCH, Achim: MowJoe: a method for automated-high throughput dissected leaf phenotyping. In: *Plant methods* 14 (2018), Nr. 1, S. 27
- [38] FAKTOR, Alon ; IRANI, Michal: “Clustering by Composition” – Unsupervised Discovery of Image Categories. In: *IEEE transactions on pattern analysis and machine intelligence* 36 (2013), Nr. 6, S. 1092–1106
- [39] DA FONA COSTA, Luciano ; CESAR JR, Roberto M.: *Shape classification and analysis: theory and practice*. Crc Press, 2018
- [40] FRANZ, Nico M.: Biological taxonomy and ontology development: scope and limitations. In: *Biodiversity informatics* 7 (2010), Nr. 1
- [41] FREEDMAN, David J. ; RIESENHUBER, Maximilian ; POGGIO, Tomaso ; MILLER, Earl K.: Categorical representation of visual stimuli in the primate prefrontal cortex. In: *Science* 291 (2001), Nr. 5502, S. 312–316
- [42] FRITZ, Michael A. ; ROSA, Stefanie ; SICARD, Adrien: Mechanisms underlying the environmentally induced plasticity of leaf morphology. In: *Frontiers in genetics* 9 (2018), S. 478
- [43] FU, Wen-Yuan ; TENG, Jiu-Cui ; TANG, Bing ; WANG, Qing-Qing ; YANG, Wei ; TAO, Lian ; WAN, Zheng-Jie ; WU, Kang-Yun ; TAN, Guo-Fei ; DENG, Ying: The Lobed-Leaf Phenotype in Brassica juncea Is Associated with the BjLMI1 Locus as Evidenced Using GradedPool-Seq. In: *Agronomy* 12 (2022), Nr. 11, S. 2696

-
- [44] FUKUNAGA, Keinosuke ; HOSTETLER, Larry D.: The estimation of the gradient of a density function, with applications in pattern recognition. In: *Information Theory, IEEE Transactions on* 21 (1975), Nr. 1, S. 32–40
- [45] GATI, Itamar ; TVERSKY, Amos: Weighting common and distinctive features in perceptual and conceptual judgments. In: *Cognitive Psychology* 16 (1984), Nr. 3, S. 341–370
- [46] GEHAN, Malia A. ; FAHLGREN, Noah ; ABBASI, Arash ; BERRY, Jeffrey C. ; CALLEN, Steven T. ; CHAVEZ, Leonardo ; DOUST, Andrew N. ; FELDMAN, Max J. ; GILBERT, Kerrigan B. ; HODGE, John G. [u. a.]: PlantCV v2: Image analysis software for high-throughput plant phenotyping. In: *PeerJ* 5 (2017), S. e4088
- [47] GIESEN, Joachim ; CAZALS, Frédéric ; PAULY, Mark ; ZOMORODIAN, Afra: The conformal alpha shape filtration. In: *The Visual Computer* 22 (2006), S. 531–540
- [48] GOËAU, Hervé ; JOLY, Alexis ; BONNET, Pierre ; SELMI, Souheil ; MOLINO, Jean-François ; BARTHÉLÉMY, Daniel ; BOUJEMAA, Nozha: Lifeclef plant identification task 2014. In: *CLEF2014 Working Notes. Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014* CEUR-WS, 2014, S. 598–615
- [49] GOLDSTONE, Robert L. ; KERSTEN, Alan ; CARVALHO, Paulo F.: Concepts and categorization. (2013)
- [50] GOMES, Jonas ; VELHO, Luiz: *From fourier analysis to wavelets*. Bd. 3. New York, USA : Springer, 2015
- [51] GONZALEZ, R. ; WOODS, R.: *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc., 2006
- [52] GRAUMAN, Kristen ; DARRELL, Trevor: Unsupervised learning of categories from sets of partially matching image features. In: *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on* Bd. 1 IEEE, 2006, S. 19–25
- [53] GRAY, Asa: *Manual of the botany of the northern United States*. USA : Ivison & Company, 1867
- [54] GUPTA, Sonal ; ROSENTHAL, David M. ; STINCHCOMBE, John R. ; BAUCOM, Regina S.: The remarkable morphological diversity of leaf shape in sweet potato (*Ipomoea batatas*): The influence of genetics, environment, and $G \times E$. In: *New Phytologist* 225 (2020), Nr. 5, S. 2183–2195
- [55] HAN, Junwei ; QUAN, Rong ; ZHANG, Dingwen ; NIE, Feiping: Robust object co-segmentation using background prior. In: *IEEE Transactions on Image Processing* 27 (2017), Nr. 4, S. 1639–1651

- [56] HAN, Kai ; REBUFFI, Sylvestre-Alvise ; EHRHARDT, Sebastien ; VEDALDI, Andrea ; ZISSERMAN, Andrew: Autonovel: Automatically discovering and learning novel visual categories. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2021), Nr. 10, S. 6767–6781
- [57] HAWKINS, WG ; LEICHNER, PK ; YANG, N-C: The circular harmonic transform for SPECT reconstruction and boundary conditions on the Fourier transform of the sinogram. In: *IEEE transactions on medical imaging* 7 (1988), Nr. 2, S. 135–138
- [58] HAWTHORNE, William ; LAWRENCE, Anna: *Plant identification: creating user-friendly field guides for biodiversity management*. Routledge, 2013
- [59] HE, Nianpeng ; LIU, Congcong ; TIAN, Miao ; LI, Meiling ; YANG, Hao ; YU, Guirui ; GUO, Dali ; SMITH, Melinda D. ; YU, Qiang ; HOU, Jihua: Variation in leaf anatomical traits from tropical to cold-temperate forests and linkage to ecosystem functions. In: *Functional ecology* 32 (2018), Nr. 1, S. 10–19
- [60] HICKEY, Michael ; KING, Clive: *The Cambridge illustrated glossary of botanical terms*. Cambridge, UK : Cambridge University Press, 2000
- [61] HOLLAND, Ian ; DAVIES, Jamie A.: Automation in the life science research laboratory. In: *Frontiers in Bioengineering and Biotechnology* 8 (2020), S. 571777
- [62] HUANG, Fei ; GAN, Yangjing ; ZHANG, Dongdong ; DENG, Fei ; PENG, Jing: Leaf shape variation and its correlation to phenotypic traits of Soybean in northeast China. In: *Proceedings of the 2018 6th International Conference on Bioinformatics and Computational Biology*, 2018, S. 40–45
- [63] JIA, Xuhui ; HAN, Kai ; ZHU, Yukun ; GREEN, Bradley: Joint representation learning and novel category discovery on single-and multi-modal data. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, S. 610–619
- [64] KANT, Immanuel: Critique of pure reason. 1781. In: *Modern Classical Philosophers, Cambridge, MA: Houghton Mifflin* (1908), S. 370–456
- [65] KASSAMBARA, Alboukadel: *Practical guide to cluster analysis in R: Unsupervised machine learning*. Bd. 1. USA : Sthda, 2017
- [66] KEENEY, Elizabeth: *The botanizers: amateur scientists in nineteenth-century America*. Chapel Hill, North Carolina, USA : Univ of North Carolina Press, 1992
- [67] KRUSCHKE, John K.: Models of categorization. In: *The Cambridge handbook of computational psychology* (2008), S. 267–301

- [68] KUHLMAN, Frank P. ; GIARDINA, Charles R.: Elliptic Fourier features of a closed contour. In: *Computer graphics and image processing* 18 (1982), Nr. 3, S. 236–258
- [69] LEE, Yong J. ; GRAUMAN, Kristen: Shape discovery from unlabeled image collections. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition* IEEE, 2009, S. 2254–2261
- [70] LETOUZEY, Rene [u. a.]: *Manual of forest botany. Tropical Africa. Vol. 2 A. Families (1st part). Vol. 2 B. Families (2nd part)*. Nogent-sur-Marne, France : Centre technique forestier tropical, 1986
- [71] LI, Mao ; AN, Hong ; ANGELOVICI, Ruthie ; BAGAZA, Clement ; BATUSHANSKY, Albert ; CLARK, Lynn ; CONEVA, Viktoriya ; DONOGHUE, Michael ; EDWARDS, Erika ; FAJARDO, Tao [u. a.]: Topological data analysis as a morphometric method: using persistent homology to demarcate a leaf morphospace. In: *Frontiers in Plant Science* 9 (2018), S. 553
- [72] LI, Mao ; FRANK, Margaret H. ; CONEVA, Viktoriya ; MIO, Washington ; CHITWOOD, Daniel H. ; TOPP, Christopher N.: The persistent homology mathematical framework provides enhanced genotype-to-phenotype associations for plant morphology. In: *Plant physiology* 177 (2017), Nr. 4, S. 1382–1395
- [73] LI, Mao ; FRANK, Margaret H. ; CONEVA, Viktoriya ; MIO, Washington ; CHITWOOD, Daniel H. ; TOPP, Christopher N.: The persistent homology mathematical framework provides enhanced genotype-to-phenotype associations for plant morphology. In: *Plant physiology* 177 (2018), Nr. 4, S. 1382–1395
- [74] LIANTONI, Febri ; PRAKISYA, Nurcahya Pradana T. ; ARISTYAGAMA, Yusfia H. ; HATTA, Puspanda: Comparative analysis of hierarchical clustering with improve feature for herbs leaves. In: *Journal of Physics: Conference Series* Bd. 1808 IOP Publishing, 2021, S. 012025
- [75] LIN, Wei-Chao ; TSAI, Chih-Fong ; HU, Ya-Han ; JHANG, Jing-Shang: Clustering-based undersampling in class-imbalanced data. In: *Information Sciences* 409 (2017), S. 17–26
- [76] LIPTON, Zachary C.: The mythos of model interpretability. In: *arXiv preprint* (2016). – <https://arxiv.org/abs/1606.03490>
- [77] LIU, Congcong ; LI, Ying ; XU, Li ; CHEN, Zhi ; HE, Nianpeng: Variation in leaf morphological, stomatal, and anatomical traits and their relationships in temperate and subtropical forests. In: *Scientific reports* 9 (2019), Nr. 1, S. 5803

- [78] LIU, Yu ; TUYTELAARS, Tinne: Residual tuning: Toward novel category discovery without labels. In: *IEEE Transactions on Neural Networks and Learning Systems* (2022), S. 1–15
- [79] LUCAS, Tim C.: A translucent box: interpretable machine learning in ecology. In: *Ecological Monographs* 90 (2020), Nr. 4, S. e01422
- [80] MÄHLER, Niklas ; SCHIFFTHALER, Bastian ; ROBINSON, Kathryn M. ; TEREBIENIEC, Barbara K. ; VUČAK, Matej ; MANNAPPERUMA, Chanaka ; BAILEY, Mark E. ; JANS-SON, Stefan ; HVIDSTEN, Torgeir R. ; STREET, Nathaniel R.: Leaf shape in *Populus tremula* is a complex, omnigenic trait. In: *Ecology and evolution* 10 (2020), Nr. 21, S. 11922–11940
- [81] MARCINKEVIČS, Ričards ; VOGT, Julia E.: Interpretability and explainability: A machine learning zoo mini-tour. In: *arXiv preprint arXiv:2012.01805* (2020)
- [82] MARK, de B. ; OTFRIED, Cheong ; MARC, van K. ; MARK, Overmars: *Computational geometry algorithms and applications*. Germany : Springer, 2008
- [83] MAZZOCCHI, Fulvio: Complexity in biology: exceeding the limits of reductionism and determinism using complexity theory. In: *EMBO reports* 9 (2008), Nr. 1, S. 10–14
- [84] MIN, Erxue ; GUO, Xifeng ; LIU, Qiang ; ZHANG, Gen ; CUI, Jianjing ; LONG, Jun: A survey of clustering with deep learning: From the perspective of network architecture. In: *IEEE Access* 6 (2018), S. 39501–39514
- [85] MOHTASHAMIAN, Mojgansadat ; KARIMIAN, Mahmood ; MOOLA, Faisal ; KAVOUSI, Kaveh ; MASOUDI-NEJAD, Ali: Automated Plant Species Identification Using Leaf Shape-Based Classification Techniques: A Case Study on Iranian Maples. In: *Iranian Journal of Science and Technology, Transactions of Electrical Engineering* 45 (2021), Nr. 3, S. 1051–1061
- [86] MOLNAR, Christoph: *Interpretable machine learning*. USA : Lulu.com, 2020
- [87] MONTAVON, Grégoire ; KAUFFMANN, Jacob ; SAMEK, Wojciech ; MÜLLER, Klaus-Robert: Explaining the predictions of unsupervised learning models. In: *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* Springer, 2022, S. 117–138
- [88] MUTALIB, Sofianita ; HASBULLAH, Nur H. ; ABDUL-RAHMAN, Shuzlina ; SHAMSUD-DIN, Mohd R. ; AB MALIK, Ariff M.: Herbal Plant Analysis Based on Leaf Features using K-Means Clustering. In: *IOP Conference Series: Earth and Environmental Science* Bd. 1019 IOP Publishing, 2022, S. 012026

- [89] NASIBOV, Efendi N. ; ULUTAGAY, Gözde: Robustness of density-based clustering methods with various neighborhood relations. In: *Fuzzy Sets and Systems* 160 (2009), Nr. 24, S. 3601–3615
- [90] NICOTRA, Adrienne B. ; LEIGH, Andrea ; BOYCE, C K. ; JONES, Cynthia S. ; NIKLAS, Karl J. ; ROYER, Dana L. ; TSUKAYA, Hirokazu: The evolution and functional significance of leaf shape in the angiosperms. In: *Functional Plant Biology* 38 (2011), Nr. 7, S. 535–552
- [91] NIKLAS, Karl J.: *Plant biomechanics: an engineering approach to plant form and function*. Chicago, IL : University of Chicago press, 1992
- [92] OLIVARES, Leonel ; VICTORINO, Jorge ; GÓMEZ, Francisco: Automatic leaf shape category discovery. In: *Pattern Recognition (ICPR), 2016 23rd International Conference on IEEE*, 2016, S. 1023–1028
- [93] OTSU, Nobuyuki: A threshold selection method from gray-level histograms. In: *Automatica* 11 (1975), Nr. 285-296, S. 23–27
- [94] OZA, Kavi K. ; DESAI, Rinku J. ; RAOLE, Vinay M.: Digital Morphometrics: A Tool for Leaf Morpho-Taxonomical Studies. In: *Indian Journal of Advanced Botany* 1 (2021), Nr. 2, S. 1–7
- [95] DE LA PAZ POLLICELLI, Maria ; IDASZKIN, Yanina L. ; GONZALEZ-JOSÉ, Rolando ; MÁRQUEZ, Federico: Leaf shape variation as a potential biomarker of soil pollution. In: *Ecotoxicology and environmental safety* 164 (2018), S. 69–74
- [96] PINAYA, Walter H. ; TUDOSIU, Petru-Daniel ; GRAY, Robert ; REES, Geraint ; NACHEV, Parashkev ; OURSELIN, Sebastien ; CARDOSO, M J.: Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. In: *Medical Image Analysis* 79 (2022), S. 102475
- [97] RADFORD, Albert E. ; AHLES, Harry E. ; BELL, C R.: *Manual of the vascular flora of the Carolinas*. Chapel Hill, North Carolina, USA : Univ of North Carolina Press, 2010
- [98] RANGANATHAN, Shiyali R.: *Prolegomena to library classification*. Madras Library Association, Madras, 1937
- [99] REEDS, Karen M.: Renaissance humanism and botany. In: *Annals of Science* 33 (1976), Nr. 6, S. 519–542
- [100] SALO, Heikki M. ; NGUYEN, Nga ; ALAKÄRPPÄ, Emmi ; KLAVINS, Linards ; HYKKERUD, Anne L. ; KARPPINEN, Katja ; JAAKOLA, Laura ; KLAVINS, Maris ;

- HÄGGMAN, Hely: Authentication of berries and berry-based food products. In: *Comprehensive Reviews in Food Science and Food Safety* 20 (2021), Nr. 5, S. 5197–5225
- [101] SAXENA, Amit ; PRASAD, Mukesh ; GUPTA, Akshansh ; BHARILL, Neha ; PATEL, Om P. ; TIWARI, Aruna ; ER, Meng J. ; DING, Weiping ; LIN, Chin-Teng: A review of clustering techniques and developments. In: *Neurocomputing* 267 (2017), S. 664–681
- [102] SHEN, Jiaming ; HAN, Jiawei: *Automated taxonomy discovery and exploration*. Springer Nature, 2022
- [103] SHIMSHONI, Ilan ; GEORGESCU, Bogdan ; MEER, Peter: 1 Adaptive Mean Shift Based Clustering in High Dimensions. In: *Nearest-neighbor methods in learning and vision: theory and practice* (2006), S. 203–220
- [104] SIFRE, Laurent ; MALLAT, Stéphane: Rotation, scaling and deformation invariant scattering for texture discrimination. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, S. 1233–1240
- [105] SMITH, David R. ; BROWN, Jennifer A.: Biological Populations. In: *Sampling rare or elusive species: concepts, designs, and techniques for estimating population parameters* (2004), S. 77
- [106] STOJNIĆ, Srđan ; VISCOSI, Vincenzo ; MARKOVIĆ, Milena ; IVANKOVIĆ, Mladen ; ORLOVIĆ, Saša ; TOGNETTI, Roberto ; COCOZZA, Claudia ; VASIĆ, Verica ; LOY, Anna: Spatial patterns of leaf shape variation in European beech (*Fagus sylvatica* L.) provenances. In: *Trees* 36 (2022), Nr. 1, S. 497–511
- [107] STROUD, Sebastian ; FENNELL, Mark ; MITCHLEY, Jonathan ; LYDON, Susannah ; PEACOCK, Julie ; BACON, Karen L.: The botanical education extinction and the fall of plant awareness. In: *Ecology and Evolution* 12 (2022), Nr. 7, S. e9019
- [108] STUESSY, Tod F.: *Plant taxonomy: the systematic evaluation of comparative data*. Columbia University Press, 2009
- [109] SUZUKI, Satoshi [u. a.]: Topological structural analysis of digitized binary images by border following. In: *Computer vision, graphics, and image processing* 30 (1985), Nr. 1, S. 32–46
- [110] THOMASSON, Amie: Categories. In: ZALTA, Edward N. (Hrsg.) ; NODELMAN, Uri (Hrsg.): *The Stanford Encyclopedia of Philosophy*. Winter 2022. Metaphysics Research Lab, Stanford University, 2022, S. 1–124
- [111] TVERSKY, Amos: Features of similarity. In: *Psychological review* 84 (1977), Nr. 4, S. 327

-
- [112] UESAKA, Y.: A new type Fourier descriptor method that is effective also to open contour. In: *IEICE Trans Inf Syst* 67 (1984), Nr. 3, S. 166–173
- [113] VAN RIJSBERGEN, Cornelis J.: Foundation of evaluation. In: *Journal of documentation* 30 (1974), Nr. 4, S. 365–373
- [114] VAZE, Sagar ; HAN, Kai ; VEDALDI, Andrea ; ZISSERMAN, Andrew: Generalized category discovery. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, S. 7492–7501
- [115] VICTORINO, J. ; GÓMEZ, F.: A comparative study of dimensionality reduction methods for p-type based contour representations. In: *Computing Colombian Conference (10CCC), 2015 10th IEEE*, 2015, S. 294–301
- [116] VICTORINO, Jorge ; GÓMEZ, Francisco: Contour analysis for interpretable leaf shape category discovery. In: *Plant methods* 15 (2019), Nr. 1, S. 1–12
- [117] VICTORINO, Jorge ; RUDAS, Jorge ; REYES, Ana M. ; PULIDO, Cristian ; CHAPARRO, Luisa F. ; ESTRADA, Camilo ; NARVAEZ, Luz A. ; GÓMEZ, Francisco: Highly Sessional Aggressive Behaviors Link to Temporal Dynamics Shared Across Space. In: *IEEE Access* 9 (2021), S. 165072–165084
- [118] VICTORINO, Jorge ; RUDAS, Jorge ; REYES, Ana M. ; PULIDO, Cristian ; CHAPARRO, Luisa F. ; NARVAEZ, Luz A. ; MARTINEZ, Darwin ; GÓMEZ, Francisco: Spatial-temporal patterns of aggressive behaviors. A case study Bogotá, Colombia. In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) IEEE*, 2020, S. 667–672
- [119] WÄLDCHEN, Jana ; MÄDER, Patrick: Plant species identification using computer vision techniques: A systematic literature review. In: *Archives of Computational Methods in Engineering* 25 (2018), Nr. 2, S. 507–543
- [120] WANG, Hui ; LIU, Pei-Liang ; LI, Jian ; YANG, Han ; LI, Qin ; CHANG, Zhao-Yang: Why more leaflets? The role of natural selection in shaping the spatial pattern of leaf-shape variation in *Oxytropis diversifolia* (Fabaceae) and two close relatives. In: *Frontiers in plant science* 12 (2021), S. 681962
- [121] WANG, Jingyu ; MA, Zhenyu ; NIE, Feiping ; LI, Xuelong: Progressive self-supervised clustering with novel category discovery. In: *IEEE Transactions on Cybernetics* (2021)
- [122] WANG, Na ; PALMROTH, Sari ; MAIER, Christopher A. ; DOMECH, Jean-Christophe ; OREN, Ram: Anatomical changes with needle length are correlated with leaf structural and physiological traits across five *Pinus* species. In: *Plant, cell & environment* 42 (2019), Nr. 1, S. 1690–1704

- [123] WASSERMAN, Larry: Topological data analysis. In: *Annual Review of Statistics and Its Application* 5 (2018), S. 501–532
- [124] WKEGLARCZYK, Stanislaw: Kernel density estimation and its application. In: *ITM Web of Conferences* Bd. 23 EDP Sciences, 2018, S. 00037
- [125] XIA, Xuan ; PAN, Xizhou ; LI, Nan ; HE, Xing ; MA, Lin ; ZHANG, Xiaoguang ; DING, Ning: GAN-based anomaly detection: A review. In: *Neurocomputing* 493 (2022), S. 497–535
- [126] YANG, Chengzhuan: Plant leaf recognition by integrating shape and texture features. In: *Pattern Recognition* 112 (2021), S. 107809
- [127] YANG, Kaiyu ; WU, Jianghao ; LI, Xinman ; PANG, Xinbo ; YUAN, Yangchen ; QI, Guohui ; YANG, Minsheng: Intraspecific leaf morphological variation in *Quercus dentata* Thunb.: a comparison of traditional and geometric morphometric methods, a pilot study. In: *Journal of Forestry Research* (2022), S. 1–14
- [128] YOUSEFI, Ehsan ; BALEGHI, Yasser ; SAKHAEI, Sayed M.: Rotation invariant wavelet descriptors, a new set of features to enhance plant leaves classification. In: *Computers and Electronics in Agriculture* 140 (2017), S. 70–76
- [129] ZHANG, Dingwen ; HAN, Junwei ; ZHAO, Long ; MENG, Deyu: Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework. In: *International Journal of Computer Vision* 127 (2019), Nr. 4, S. 363–380
- [130] ZHANG, Quan-shi ; ZHU, Song-Chun: Visual interpretability for deep learning: a survey. In: *Frontiers of Information Technology & Electronic Engineering* 19 (2018), Nr. 1, S. 27–39
- [131] ZHENG, Yuanting ; XU, Fei ; LI, Qikai ; WANG, Gangjun ; LIU, Na ; GONG, Yaming ; LI, Lulu ; CHEN, Zhong-Hua ; XU, Shengchun: QTL mapping combined with bulked segregant analysis identify SNP markers linked to leaf shape traits in *Pisum sativum* using SLAF sequencing. In: *Frontiers in genetics* 9 (2018), S. 615
- [132] ZVEREVA, Elena L. ; KOZLOV, Mikhail V.: Biases in ecological research: attitudes of scientists and ways of control. In: *Scientific Reports* 11 (2021), Nr. 1, S. 226