ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2023 Управление, вычислительная техника и информатика № 63

Tomsk State University Journal of Control and Computer Science

Original article
doi: 10.17223/19988605/63/13

Application of *k*-means clustering and histogram analysis to automate preprocessing of images of discomycetes obtained in the habitat

Darya A. Filimonova[1], Irina G. Vorob'eva[2], Alexander Yu. Filimonov[3]

[1, 2] Novosibirsk State Pedagogical University, Novosibirsk, Russian Federation
[2] Central Siberian Botanical Garden of the Siberian Branch of the RAS, Novosibirsk, Russian Federation
[3] Ural Federal University named after the first President of Russia, Boris Yeltsin, Yekaterinburg, Russian Federation
[1] darya.filimonova@gmail.com
[2] vorobig@ngs.ru
[3] a.filimonov@urfu.ru

Abstract. The study of biological diversity requires a thorough inventory of all groups of organisms, including destructors, among which fungi play a significant role. Discomycetes, a group of orders of fungi of the Ascomycota phylum, require close attention from researchers due to their low level of knowledge. The paper proposes an approach to automating the process of inventory of representatives of this group of orders and presents a prototype of a software package that allows one to identify the presence of fruit bodies of discomycetes in photographs taken in the natural habitat. A feature of the proposed solution is the application of the *k*-means clustering method, the use of scaled histograms to determine the presence of an image of the fruit body of Discomycetes in this image, and the prospects for using this tool in machine learning are described using neural networks.

Keywords: clustering; computer vision; machine learning, discomycetes; biodiversity.

For citation: Filimonova, D.A., Vorob'eva, I.G., Filimonov, A.Yu. (2023) Application of *k*-means clustering and histogram analysis to automate preprocessing of images of discomycetes obtained in the habitat. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika – Tomsk State University Journal of Control and Computer Science*. 63. pp. 111–117. doi: 10.17223/19988605/63/13

Научная статья
УДК 58.002; 004.932
doi: 10.17223/19988605/63/13

Применение кластеризации *k*-means и анализа гистограмм для автоматизации предварительной обработки изображений дискомицетов, полученных в среде обитания

Дарья Александровна Филимонова[1], Ирина Геннадьевна Воробьева[2], Александр Юрьевич Филимонов[3]

[1, 2] Новосибирский государственный педагогический университет, Новосибирск, Российская Федерация
[2] Центральный сибирский ботанический сад СО РАН, Новосибирск, Российская Федерация
[3] Уральский федеральный университет им. первого президента России Б.Н. Ельцина, Екатеринбург, Российская Федерация
[1] darya.filimonova@gmail.com
[2] vorobig@ngs.ru
[3] a.filimonov@urfu.ru

Аннотация. Изучение биологического разнообразия требует проведения тщательной инвентаризации всех групп организмов, в том числе деструкторов, среди которых значительную роль играют грибы. Дискомицеты – группа порядков грибов отдела Ascomycota – требуют пристального внимания со стороны исследователей ввиду недостаточной изученности. В работе предлагается подход к автоматизации процесса инвентаризации представителей данной группы порядков и представлен прототип программного комплекса, позволяющий выявить наличие плодовых тел дискомицетов на фотографиях, сделанных в естественной среде обитания. Особенностью предлагаемого решения являются применение метода кластеризации k-means, использование масштабированных гистограмм для определения наличия образа плодового тела дискомицета на данном изображении. Также описаны перспективы использования данного средства в машинном обучении с применением нейросетей.

Ключевые слова: кластеризация; компьютерное зрение; машинное обучение; дискомицеты; биоразнообразие.

***Для цитирования*:** Филимонова Д.А., Воробьева И.Г., Филимонов А.Ю. Применение кластеризации k-means и анализа гистограмм для автоматизации предварительной обработки изображений дискомицетов, полученных в среде обитания // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2023. № 63. С. 111–117. doi: 10.17223/19988605/63/13

Introduction

Fungi are one of the oldest groups of organisms that play an important role in the circulation of organic substances. Thus, fungi that enter into symbiosis with plants are of high ecological importance, helping the latter to obtain inaccessible mineral nutrition resources. In particular, it is thanks to mycorrhizal fungi that phosphorus is incorporated from the geological cycle into the biological one [1]. In addition, as destructors, fungi are of exceptional importance in the biological cycle of forest ecosystems, since they have the necessary and self-sufficient enzyme systems that allow them to carry out a complete biochemical conversion of wood compounds [1–3]. The destructors of organic matter also include discomycetes, a group of orders of fungi of the Ascomycota phylum. Most of them are saprotrophs, developing on soil, litter, wood, excrement, places of fire [4]. From a biological point of view, discomycetes have not been studied sufficiently, and the study of this group is accompanied by difficulties, including high synonymy, as well as the lack of parity between phenotypic and genetic methods of determination.

Work on the conservation and restoration of biological diversity, as well as on the introduction of this group of organisms into biotechnological processes, is impossible without an inventory of flora and a detailed study of the spread of discomycetes in natural and anthropogenic communities. Currently, most researchers use a methodology according to which the initial determination of the collected samples is carried out according to phenotypic characteristics using a magnifying glass or a light microscope with various magnifications. Further identification includes the use of micropreparations [5–8]. In addition, recently, along with more familiar methods of determining the systematic position of fungi, a relatively new technique has been used, that of machine vision. The use of this method in identifying mushrooms has several problems. The constant development of the taxonomy of fungi leads to a continuous stream of reclassifications and the introduction of new names, which causes difficulties in the taxonomic differentiation of fungi [9]. The updated classification of representatives of this kingdom is increasingly based not on external signs but on the analysis of cellular structure and molecular genetic studies. Despite the fact that there is a need to constantly compare the information obtained by molecular methods with the data of previous studies conducted by traditional methods using morphological features, new knowledge allows us to shed light on the existing problems in the classification of fungi.

The paper considers proposals for the organization of a software prototype for preprocessing images of discomycetes obtained in the habitat. During this processing, images of discomycetes, often growing in groups, should be separated from each other, from the substrate on which they grow, and from the herbaceous cover. The selected fragments of the original image are further processed for subsequent identification and inventory.

The paper presents comparative results of the implementation of this algorithm using various color models of the original image and confirms the conclusion that the prospects of using non-additive color models to solve the problem [10]. Subsequent processing of image fragments was carried out based on the analysis of histograms of the distribution of color components, which is traditionally used for similar purposes [11]. The results presented in the paper show that the application of the described methods provides reliable determination of the presence of images of fruit bodies of discomycetes in fragments of the original image and automated generation of reporting data on observation. The data obtained also allows us to make an assumption about the possibility of using histograms of the color distribution of image fragments as key features for automating the process of identifying discomycetes using machine learning tools.

1. Materials and methods

The objects of the study were the fruit bodies of discomycetes, images of which were obtained in habitat.

The prototype of the software package was developed in Python; computer vision programming requires representations of vectors and matrices and operations on them. This is handled by the NumPy Python module, where both vectors and matrices are represented as an array. NumPy is a widely used package for scientific computing using Python. It contains a number of useful concepts, such as arrays (for representing vectors, matrices, images, and more) and linear algebra functions. Arrays allow you to perform important operations, such as matrix multiplication, transposition, solving systems of equations, vector multiplication, and normalization, which are necessary, in particular, for image processing. To visualize the results, the Matplotlib module was used, which has open source code and creates high-quality illustrations. OpenCV is a C++ library with modules that cover many areas of computer vision. OpenCV has functions for reading and writing images, as well as for matrix operations and mathematical libraries [12].

Image preprocessing includes Gaussian blur, which simplifies contour detection by reducing the noise level in the image; clustering by k-means; and contouring.

Clustering divides an image into its constituent regions or objects. The degree of detail to which such a separation is brought depends on the task being solved. In other words, clustering should be stopped when the objects or areas of interest are detected. Clustering images that are not trivial is one of the most difficult image processing tasks. The ultimate success of computer image analysis procedures is largely determined by the accuracy of clustering. For this reason, considerable attention should be paid to improving its reliability [13]. Clustering can be used both for image recognition and for splitting an image into fragments for subsequent processing, etc. [12, 14].

When clustering an image using the k-means method, each of its color points is assigned to one of the clusters with the nearest average color value of the centroid. The k-means algorithm is an iterative procedure that sequentially refines the average values of clusters until convergence is achieved [13]. The only significant disadvantage of the chosen algorithm is the tendency to align the cluster sizes. Indeed, since the inertia criterion is determined by the sum of the minimum values of the root-mean-square deviations of the cluster point coordinates from the centroids [15], it is more profitable for the algorithm to "stretch" clusters, which, with a small value of their number, can significantly affect the clustering result.

This disadvantage will be especially pronounced in cases where the values of the color coordinates of the original image tend to have a normal distribution. That is why, as studies show, the results of the clustering of color images differ when using different color models [10]. In this case, images with the HSL color model were used for clustering, which allowed us to achieve better results.

The images were contoured using the canny algorithm. Further processing was performed using the getStructuringElement() and morphologyEx() functions. The obtained areas of the image, which with a certain degree of probability contained images of the fruit bodies of mushrooms, were analyzed using histograms.

Histograms are the basis for numerous image processing methods; histogram-based data analysis is one of the most popular solutions for many image processing tasks, such as object recognition and classification. Studies show that the generalized information contained in the histogram is very useful for solving problems such as image compression and segmentation [13, 16].

A raw histogram is an integral characteristic of an image that shows only the number of points depending on their color, and therefore is not a convenient tool for studying images of complex objects. However, when analyzing images of such simple objects as discomycetes, for the classification of which one of the key parameters is the predominant color, they can be useful. To ensure the commensurability of histograms of image fragments obtained as a result of clustering and contouring, they were reduced to an identical dimension along the abscissa axis corresponding to the color tone. This action made it possible to bring the histograms of all fragments of the image to a unified view. The next stage was normalization, which took place as follows: the entire array of values of the color vectors of the histogram was divided into the total sum of the vectors, so that sections of the original image with different areas became comparable. Thus, after processing, the histograms were an array of data divided into 128 groups, each containing two consecutive shades to reduce the dimension of the resulting histograms.

The reliability of the assumption about the relationship between the appearance of histograms and the presence or absence of an image of fruit bodies of mushrooms in a particular area of the image was checked using a correlation coefficient table calculated using the corresponding functions of the Pandas package. The correlation was displayed and analyzed using the study of the resulting heat map.

The final stage of the development of the prototype of the software was the use of a neural network to automate the determination of the presence of fruit bodies on image fragments. The data for training the neural network was prepared as follows: scaled normalized histograms of even fragments of the original image were used as a training set, and odd fragments were used as a test set. A simple sequential fully connected neural network was created, the input fully connected layer of which consisted of 128 neurons with 127 inputs due to the dimension of the histogram; the output consisted of 2 neurons. The choice of exactly this number of output neurons is due to the prospect of using a similar scheme for the classification of discomycetes. The Relu method was chosen to activate the input neurons, and the Softmax method was used in the output layer. Cross Entropy was used as a loss function, and Adam was used as a learning optimizer.

2. Results and discussion

The initial image of the fruit bodies of discomycetes and the result of the application of k-means are shown in Fig. 1. As can be seen in the image, a leaf litter similar in color was also attributed to a cluster containing images of the fruit bodies of discomycetes.

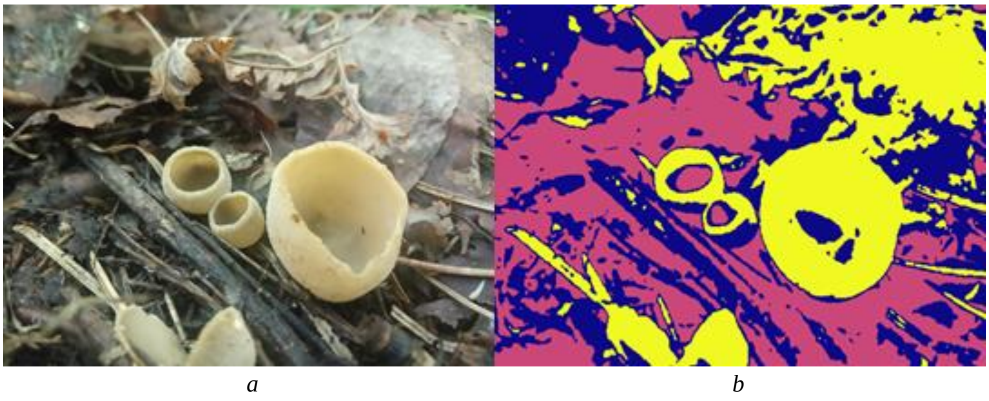

Fig. 1. Fruit bodies of fungi of the order Pezizales on the substrate:
a) the original image; *b*) the result of clustering

The result of image contouring is shown in Fig. 2. Rough contours are formed, both containing (fragment numbers 0, 6) and not containing images of the fruit bodies of discomycetes.

During the study, several photographs obtained in the habitat were processed, containing up to 60 fragments obtained as a result of clustering. At the same time, mushrooms contained no more than 10% of the formed fragments. Thus, the task of automating the determination of whether this fragment contains an image of the fruit body of a mushroom or not has become urgent. The assumption that this can be done based on the form of a scaled normalized histogram was verified by calculating the correlation matrix of histograms of image fragments presented in the form of a heat map (Fig. 3).

The light fragments of this matrix obtained using the Pearson linear correlation coefficient correspond to high values of the coefficient, the dark ones correspond to low values. The highest correlation values are fragments of the original image, which depict similar objects: fruit bodies of discomycetes, leaf litter, etc.

Fragments of plots containing images of fungi correlate with each other (the correlation coefficient is 0.797). This confirms the assumption that it is possible to determine the presence of an image of the fruit body of a mushroom on a fragment of the original image using the form of a histogram of the distribution of the color tone.

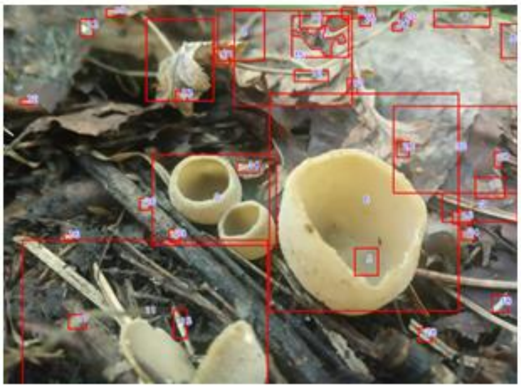

Fig. 2. An image containing the contours of the fruit bodies of discomycetes

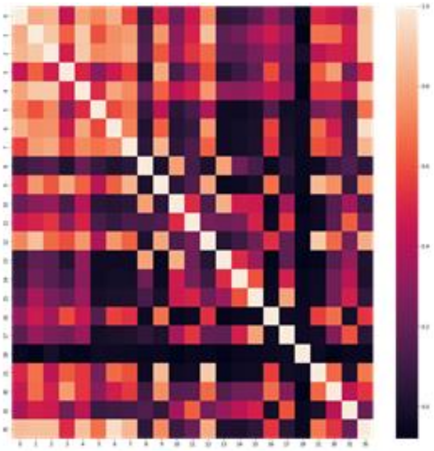

Fig. 3. Correlation matrix calculated from the contours of histograms of image fragments after cleaning

The final stage was the creation of a neural network. The neural network was trained on 50 fragments of the original images. The training included 150 epochs. At 65 iterations of the training, the accuracy value stabilized at 90% (Fig. 4).

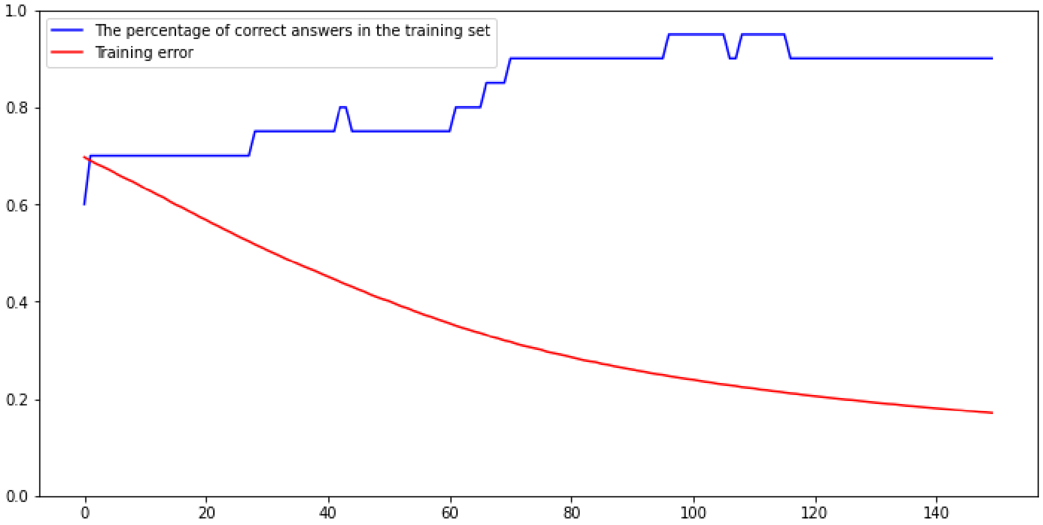

Fig. 4. The result of training a neural network prototype

Despite the extremely limited size of the training set, testing on the test set gave results of 85 to 88%. This allows us to speculate on the possibility of not only automating the process of detecting fruit bodies of discomycetes in photographs taken in natural conditions, but also, with a large amount of data, automating the process of classifying them based on phenotypic characteristics.

Conclusion

The paper proposes an approach to automating the process of inventory of representatives of discomycetes and presents the results of using prototypes of programs of this technological cycle, confirming the possibility of detecting the presence of fruit bodies of discomycetes in photographs taken in the natural habitat.

The proposed solution included the application of the k-means clustering method, the use of normalized scaled histograms of image fragments obtained after clustering to determine the presence of an image of the fruit body of representatives of the Discomycete order group on it, as well as the study of the possibility of automating this detection process using a neural network prototype.

Thus, the obtained results confirmed the hypothesis that normalized scaled histograms of the color distribution of image fragments can be used as key features for determining the presence of discomycetes in the image. In addition, an assumption was made about the possibility of automating both the identification process and, in the future, the classification of discomycetes using machine learning tools with the accumulation of a sufficient amount of material.

References

1. Mukhin, V.A. (1999) Griby i ikh rol' v prirode i razvitii tsivilizatsii [Mushrooms and their role in nature and the development of civilization] *Izvestiya Ural'skogo gosudarstvennogo universiteta – Proceedings of the Ural State University*. 12. pp. 64–69.
2. Bogacheva, A.V. (2018) New and interesting finds of discomycetes in the territory of Khabarovsk region. *Biota i sreda zapovednykh territoriy – Biota and Environment of Natural Areas*. 2. pp. 41–53.
3. Filippova, N., Arefiev, S., Zvyagina, E., Kapitonov, V., Makarova, T., Mukhin, V. & Paukov, A. (2020) Fungal literature records database of the Northern West Siberia (Russia). *Biodiversity Data Journal*. 8. DOI: 10.3897/bdj.8.e52963
4. Smitskaya, M.F. (1980) *Flora gribov Ukrainy: Operkulyatnye diskomitsety* [Mushroom flora of Ukraine: Operculate discomycetes]. Kyiv: Naukova dumka.
5. Prokhorov, V.P. (2004) *Opredelitel' gribov Rossii. Diskomitsety* [Identificator of Russian mushrooms. Discomycetes]. Vol. 1. Moscow: Association of Scientific Publications of the CMC.
6. Popov, E.S. (2005) *Diskomitsety Severo-Zapada evropeyskoy chasti Rossii: Leningradskaya, Novgorodskaya, Pskovskaya oblasti, g. Sankt-Peterburg* [Discomycetes of the North-West of the European part of Russia (Leningrad, Novgorod, Pskov regions, St. Petersburg)]. PhD thesis. St. Petersburg.
7. Bogacheva, A.V. (1997) *Diskomitsety zapovednikov Primorskogo kraya* [Discomycetes in the nature reserves of Primorsky Krai]. PhD Thesis. Vladivostok.
8. Bogacheva, A.V. (2008) *Diskomitsety (Ascomycota: Helotiales, Neolectales, Orbiliales, Pezizales, Thelebolales) yuga Dal'nego Vostoka Rossii* [Discomycetes (Ascomycota: Helotiales, Neolectales, Orbiliales, Pezizales, Thelebolales) of the South of the Russian Far East]. Dr. Diss. Vladivostok.
9. Vu, D., Groenewald, M. & Verkley, G. (2020). Convolutional neural networks improve fungal classification. *Scientific Reports*. 10(1). DOI: 10.1038/s41598-020-69245-y
10. Jurio, A., Pagola, M., Galar, M., Lopez-Molina, C. & Paternain, D. (2010). A Comparison Study of Different Color Spaces in Clustering Based Image Segmentation. *Communications in Computer and Information Science*. pp. 532–541. DOI: 10.1007/978-3-642-14058-7_55
11. Solomon, C. & Gibson, S. (2011) *Fundamentals of Digital Image Processing.* Chichester, UK: Wiley & Sons, Limited, John.
12. Erik, J. & Solem, J.E. (2012) *Programming computer vision with Python.* Beijing; Cambridge; Sebastopol [etc.]: O'Reilly Media, Inc, USA.
13. Gonzalez, R.C. & Woods, R.E. (2008) *Digital Image Processing.* Upper Saddle River, N.J: Prentice Hall.
14. Mittal, H., Pandey, A. C., Saraswat, M., Kumar, S., Pal, R. & Modwel, G. (2021). A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets. *Multimedia Tools and Applications*. 81(24). pp. 35001–35026. DOI: 10.1007/s11042-021-10594-9
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. & Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *The Journal of Machine Learning Research*. 12. pp. 2825–2830. DOI: 10.5555/1953048.2078195
16. Blachnik, M. & Laaksonen, J. (2008) Image classification by histogram features created with learning vector quantization. *Artificial Neural Networks – ICANN 2008*. (5163). pp. 827–836. DOI: 10.1007/978-3-540-87536-9_85

***Information about the authors*:**

Filimonova Darya A. (Post-graduate Student, Department of Biology and Ecology, Institute of Natural and Social and Economic Sciences, Novosibirsk State Pedagogical University, Novosibirsk, Russian Federation). E-mail: darya.filimonova@gmail.com

Vorob'eva Irina G. (Doctor of Biological Sciences, Associate Professor, Professor of the Department of Biology and Ecology of the Institute of Natural and Social and Economic Sciences of Novosibirsk State Pedagogical University; leading researcher at the Laboratory of Dendrology of the Central Siberian Botanical Garden of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russian Federation). E-mail: vorobig@ngs.ru

Filimonov Alexander Yu. (Senior lecturer of the Ural Federal University named after the first President of Russia Boris Yeltsin, Yekaterinburg, Russian Federation). E-mail: a.filimonov@urfu.ru

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

***Информация об авторах*:**

Филимонова Дарья Александровна – аспирант кафедры биологии и экологии Института естественных и социально-экономических наук Новосибирского государственного педагогического университета (Новосибирск, Россия). E-mail: darya.filimonova@gmail.com

Воробьева Ирина Геннадьевна – доктор биологических наук, доцент, профессор кафедры биологии и экологии Института естественных и социально-экономических наук Новосибирского государственного педагогического университета; ведущий научный сотрудник лаборатории дендрологии Центрального сибирского ботанического сада СО РАН (Новосибирск, Россия). E-mail: vorobig@ngs.ru

Филимонов Александр Юрьевич – старший преподаватель департамента информационных технологий и автоматики Уральского федерального университета им. первого президента России Б.Н. Ельцина (Екатеринбург, Россия). E-mail: a.filimonov@urfu.ru

Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Received 20.09.2022; accepted for publication 09.06.2023

Поступила в редакцию 20.09.2022; принята к публикации 09.06.2023