

Сегментация и отображение загрязнений водной среды на основе метода К-средних

М.Б. Хасанов¹, С.А.К. Диане²

¹ НИУ ВШЭ, ул. Мясницкая, д. 20, г. Москва, 101000, Россия

² ИППУ РАН, ул. Профсоюзная, 65, г. Москва, 117997, Россия

Аннотация

В докладе представлено исследование текущего состояния систем обнаружения загрязнений водной поверхности. Предложена формализация карты центроидов для трёхканального аэрофотоснимка. Рассмотрен пример использования алгоритма К-средних для кластеризации участков местности на тестовых аэрофотоснимках. Дана визуализация результатов кластеризации аэрофотоснимков для разного количества центроидов и результатов сегментации загрязнения. Приведена блок-схема алгоритма кластеризации, выявлены его преимущества и недостатки. Описана структурная схема программного обеспечения, разработанного на языке Python с применением кроссплатформенных библиотек компьютерной графики. Произведена оценка точности использования алгоритма кластеризации с применением метрики F1. Предварительные экспериментальные исследования показали, что включение эксперта в контур принятия решений позволяет повысить гибкость программы, благодаря возможности выделять целевую область, изменять параметры количества кластеров, точность сегментации.

Ключевые слова

Алгоритм К-средних, кластеризация, сегментация изображений, обнаружение загрязнений, F1-мера.

Segmentation and Visualization of Water Pollution Based on the K-means Method

M.B. Khasanov¹, S.A.K. Diane²

¹ HSE University, 20 Myasnitskaya str., Moscow, 101000, Russia

² ICS RAS, 65 Profsoyuznaya str., Moscow, 117997, Russia

Abstract

The paper presents a study of the current state of water pollution detection systems. A formalization of the centroid map for a three-channel aerial photograph is proposed. An example of using the K-means algorithm for clustering terrain and water areas on test aerial photographs is considered. The visualization of the results of clustering of aerial photographs for a different number of centroids is given as well as the results of pollution segmentation. A block diagram of the clustering algorithm is presented. Its advantages and disadvantages are identified. The structure of the developed software using Python and cross-platform computer graphics libraries is described. An assessment of the accuracy of using the clustering algorithm using the F1-measure is performed. Preliminary experimental studies showed that the inclusion of an expert in the contour of decision-making allows increasing the flexibility of the program, due to the possibility of selecting a target area, choosing the number of clusters and segmentation accuracy.

Keywords

K-means algorithm, clustering, image segmentation, pollution detection, F1-measure.

ГрафиКон 2023: 33-я Международная конференция по компьютерной графике и машинному зрению, 19-21 сентября 2023 г., Институт проблем управления им. В.А. Трапезникова Российской академии наук, г. Москва, Россия

EMAIL: mbkhasanov@edu.hse.ru (М.Б. Хасанов); diane1990@yandex.ru (С.А.К. Диане)

ORCID: 0000-0002-8690-6422 (С.А.К. Диане)



© 2023 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1. Введение

Проблема экологического мониторинга и, в частности, задача обнаружения загрязнений на поверхности воды приобретают все большую актуальность ввиду возрастающих темпов промышленного производства, а также приобретения и эксплуатации плавательных судов.

Исследование текущего состояния систем визуального обнаружения загрязнений на водной поверхности показало, что ключевыми подходами, используемыми на сегодняшний день, являются: корреляционный, гистограммный и нейросетевой методы.

Суть корреляционного анализа заключается в расчете коэффициентов корреляции эталонного изображения с фрагментами анализируемой фотографии. Данные коэффициенты могут принимать, как правило, положительные и отрицательные значения. Нормализованное значение коэффициента корреляции может быть интерпретировано как вероятность нахождения эталонного фрагмента в рассматриваемой позиции [1].

Гистограммный метод хорошо работает для изображения, где присутствует большое число объектов с разнообразной яркостью. Для каждого из анализируемых фрагментов изображения строится вектор признаков – гистограмма, содержащая информацию о количестве пикселей, имеющих тот или иной оттенок. Данная гистограмма затем сравнивается с эталонным вектором. Сопоставление гистограмм, с одной стороны, более вычислительно эффективно, нежели сравнение исходных растров, а с другой – более устойчиво к различного рода смещениям объектов на фотографии [2].

В последние десятилетия хорошо зарекомендовали себя нейросетевые методы. Данные методы базируются на применении различных типов нейронных сетей, в особенности сверточных. Данная технология, по сути, сочетает в себе функционал корреляционного и гистограммного подходов, позволяя извлекать ключевые характеристики или признаки наблюдаемых образов, классифицировать эти образы, а также находить приближенные решения оптимизационных задач [3].

Возможным недостатком нейросетевого подхода является сложность отладки найденных закономерностей по причине неявного представления знаний, хранимых в формате массива весовых коэффициентов, отражающих силу связей между нейронами.

Вместе с тем, наряду с перечисленными подходами перспективно использовать сегментацию изображений на базе методов кластеризации [4]. Такой подход достаточно прост в реализации, не требует подготовки обширной базы обучающих примеров и интуитивно понятен благодаря простому алгоритму визуализации.

2. Разработка алгоритма кластеризации загрязнений водной среды на основе метода K-средних

Классическим примером кластеризации является разграничение точек на двумерной плоскости. Однако в случае с анализом цифровых изображений размерность задачи существенно повышается. Анализу подлежат не только координаты пикселей аэрофотоснимков, но и их оттенки, выраженные в цветовом пространстве RGB или HSV.

2.1. Построение карты центроидов кластеризации на эталонном аэрофотоснимке с применением алгоритма K-средних

Карта центроидов кластеризации может формально быть представлена в виде множества $M = \{p_1, \dots, p_n\}$, состоящего из точек $p_i = \{id, r, g, b\}$, где id – порядковый номер кластера, r , g , b – компоненты красного, зеленого и синего каналов соответственно.

Для построения подобной карты необходимо проанализировать один или несколько тестовых наборов данных (в рассматриваемом случае – точек аэрофотоснимков). При этом центры кластеров, первоначально расположенные в случайных позициях, сместятся в местоположения, близкие к оптимальным.

Обобщенное описание алгоритма K-средних выглядит следующим образом [5]:

1. случайным образом создаются k точек, в дальнейшем будем называть их центрами кластеров;
2. для каждой точки ставится в соответствии ближайший к ней центр кластера;
3. вычисляются средние арифметические точек, принадлежащих к определённому кластеру; именно эти значения становятся новыми центрами кластеров;
4. шаги 2 и 3 повторяются до тех пор, пока пересчёт центров кластеров будет приводить к изменениям в местоположениях кластеров.

В качестве наглядного примера рассмотрим аэрофотоснимок побережья реки Волга, представленный на рисунке 1. Данная цифровая фотография содержит достаточно большое число типов земной и водной поверхности, такие как: сухой песок, мокрый песок, трава, сухая трава, прибрежная вода, вода, загрязненная трава, металл.

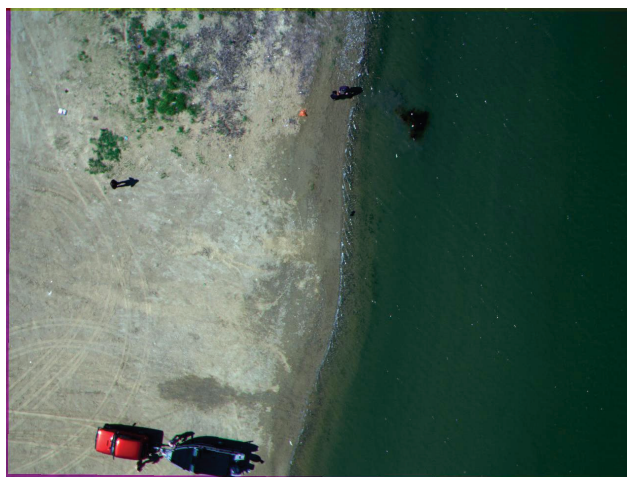


Рисунок 1 – Калибровочный аэрофотоснимок

Такое количество классов может негативно сказаться на результатах обработки аэрофотоснимка. Для упрощения процесса кластеризации целевой области произведем кадрирование и подстройку яркости фотографии (рисунок 2).

Дополнительно на обрезанной версии аэрофотоснимка пользователь задает целевую область в виде многоугольника (рисунок 2, б). Наличие такой области позволяет задать критерий оценки точности автоматической сегментации загрязнения на поверхности воды.

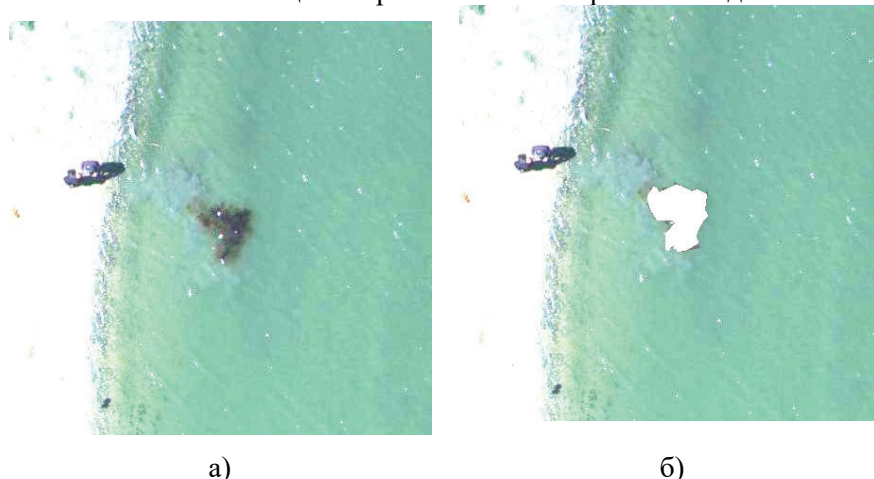


Рисунок 2 – Обработанные изображения: а) осветленный и обрезанный снимок; б) снимок с выделенной целевой областью

Оценка точности может базироваться на сопоставлении площадей эталонного многоугольника и фактически сегментированной зоны. Аналитическим критерием, пригодным для расчета подобного рода оценки является метрика F1.

Вычислить F1-меру можно по формуле:

$$F_1 = \frac{2 * P_r * R_e}{P_r + R_e},$$

где P_r – точность (precision), R_e – полнота (recall).

В свою очередь P_r и R_e можно вычислить по формулам:

$$P_r = \frac{TP}{TP + FP},$$

где TP – число истинно положительных классификаций относительно общего числа положительных наблюдений, а FP – число ложноположительных результатов относительно общего числа положительных наблюдений.

$$R_e = \frac{TP}{TP + FN},$$

где TP – число истинно положительных классификаций относительно общего числа положительных наблюдений, а FN – число ложноотрицательных результатов относительно общего числа отрицательных наблюдений.

На основе расчета оценок F1 для областей загрязнения, с применением различных параметров кластеризации, можно получить число центроидов, максимизирующее точность сегментации загрязнений.

2.2. Обобщенная структурная схема разработанного программно-алгоритмического обеспечения

Для проверки эффективности предложенного подхода было разработано программное обеспечение на языке Python, обобщенная структура которого представлена на рисунке 3.

Основными модулями в данной схеме являются:

1. процедура нахождения достаточного количества кластеров, опирающаяся на многократный пересчет алгоритма K-средних;
2. модуль сегментации загрязнений, позволяющий для выбранного количества кластеров в пределах заданной погрешности обнаружить целевую область на поверхности водоёма;
3. интерфейс пользователя, включающий как визуализацию, так и вывод численных оценок.

В качестве стандартных программных компонент для визуализации результатов были использованы кроссплатформенные библиотеки PIL, OpenCV, Matplotlib.

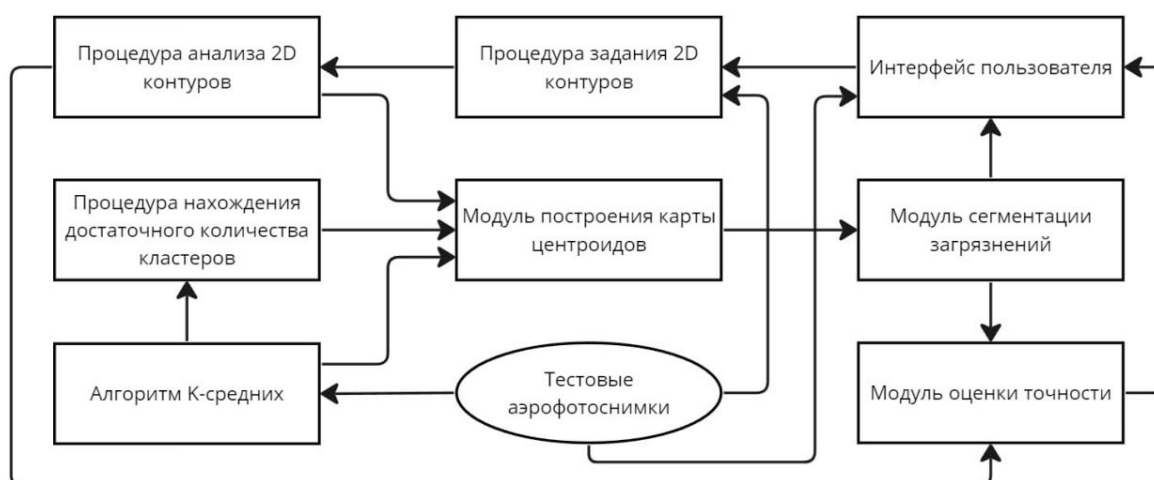


Рисунок 3 – Структурная схема программного обеспечения

Обобщенный алгоритм работы программного обеспечения для кластеризации загрязнений на поверхности водной среды представлен на рисунке 4.



Рисунок 4 – Блок-схема алгоритма

Как видно из блок-схемы, алгоритм включает ввод данных пользователем, их обработку и вывод полученных результатов.


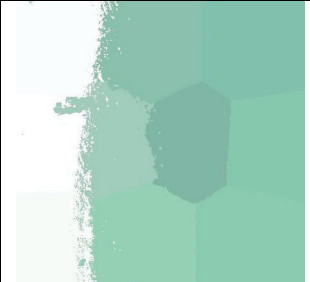
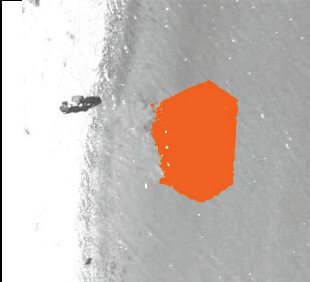


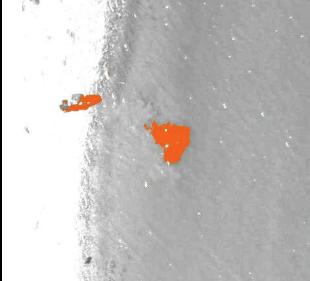


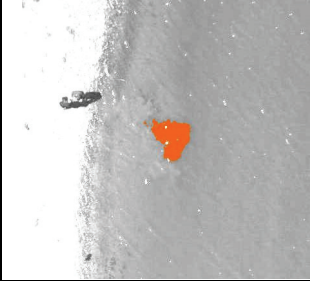

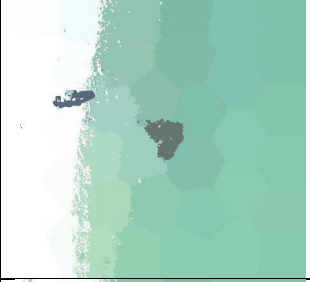
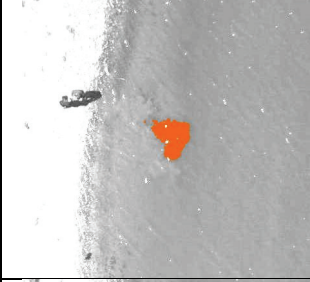

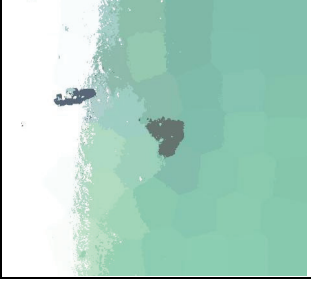
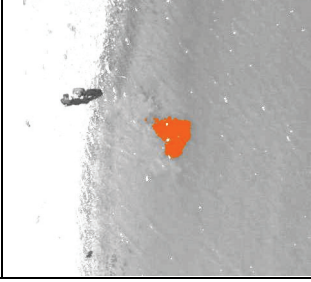
Основными действиями в данном алгоритме являются:

1. задание эталонной области, с помощью которой будет вычислена точность работы алгоритма;
2. задание количества кластеров N и погрешности E с целью наиболее оптимальной сегментации изображения;
3. применение алгоритма кластеризации K -средних с параметром N для последующей сегментации точек загрязнения;
4. сегментация точек загрязнения для наглядного отображения целевой области, выделенной экспертом.

3. Кластеризация тестовых аэрофотоснимков с применением оптимального числа кластеров

Серия экспериментов по выбору оптимального количества центроидов для алгоритма К-средних (таблица 1) показала, что величина F-меры сначала увеличивается при увеличении количества кластеров, так как разбиение на кластеры становится всё точнее и точнее.

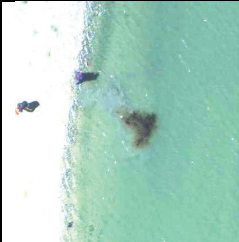

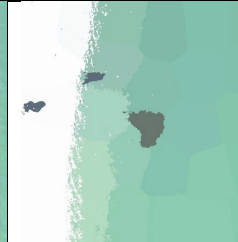
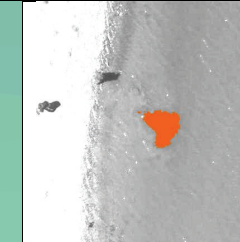


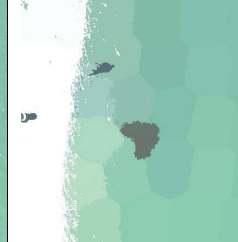
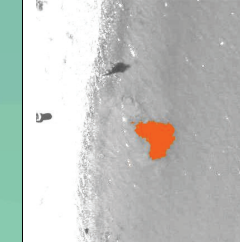


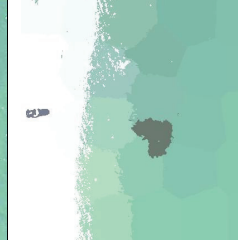
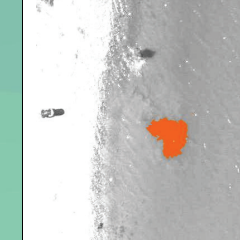



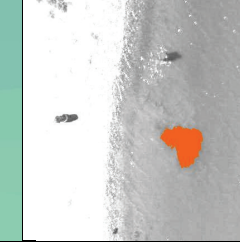
Таблица 1 – Кластеризация первого снимка

Кол-во кластеров	Изначальный снимок	Снимок после кластеризации	Выделенная область	F-мера
10				0.232
20				0.594
30				0.659
40				0.642
50				0.631

Начиная с 30 кластеров увеличение останавливается, так как достигнута необходимая точность. Примечательно, что, начиная с 40 кластеров, идет уменьшение величины F-меры, поскольку область, выделяемая при кластеризации, не включает в себя блики от воды и относит их не к загрязнению, а к обычной воде. Исходя из этого, выберем число центроидов для тестовых снимков равным 30.

Результат сегментации загрязнений на тестовых аэрофотоснимках с применением алгоритма, описанного в разделе 2.2, и выбранного количества центроидов кластеризации приведен в таблице 2.

Таблица 2 – Таблица для четырех тестовых аэрофотоснимков

№ фото	Изначальный снимок	Область, выделенная экспертом	Снимок после кластеризации	Выделенная область	F-мера
2					0.681
3					0.691
4					0.684
5					0.673

4. Заключение

Анализируя полученные результаты, можно выделить несколько особенностей применения алгоритма K-средних для обнаружения загрязнений на водной поверхности.

Главным преимуществом метода кластеризации для решения данной задачи являются наглядная визуализация и возможность последующего анализа полученных изображений, так как геометрические параметры загрязнения достаточно точно определены. Отметим, что включение эксперта в контур принятия решений позволяет повысить гибкость программы, благодаря возможности выделять целевую область, изменять параметры количества кластеров, точность сегментации.

Недостатками в использовании алгоритма кластеризации для отображения загрязнений водной среды являются:

1. сравнительно низкая скорость работы алгоритма при обработке изображений большой размерности;
2. возможность ложных срабатываний алгоритма в условиях наличия бликов на поверхности воды (этот недостаток можно устранить путем расширения вектора параметров для каждой из кластеризуемых точек);
3. необходимость ручного подбора количества кластеров (данный недостаток можно решить путем использования метода локтя [6])

Ряд из выявленных недостатков можно решить за счет использования вместо алгоритма К-средних алгоритма DBSCAN, в котором не нужно задавать количество кластеров [7].

5. СПИСОК ИСТОЧНИКОВ

- [1] Im J., Jensen J. R., Tullis J. A. Object-based change detection using correlation image analysis and image segmentation // *International Journal of Remote Sensing*. 2008. № 29(2), P. 399–423. DOI: 10.1080/01431160601075582.
- [2] Liu G. H., Yang J. Y. Deep-seated features histogram: A novel image retrieval method // *Pattern Recognition*. 2021. № 116 (1). DOI: 10.1016/j.patcog.2021.107926.
- [3] Ilesanmi A. E., Ilesanmi T. Methods for image denoising using convolutional neural network: a review // *Complex and Intelligent Systems*. 2021. № 7(5), P. 2179–2198. DOI: 10.1007/s40747-021-00428-4.
- [4] A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets / H. Mittal, A. C. Pandey, M. Saraswat, S. Kumar, R. Pal, G. Modwel // *Multimedia Tools and Applications*. 2022. № 81(24), P. 35001–35026. DOI: 10.1007/s11042-021-10594-9.
- [5] Yuryev G. A., Verkhovskaya E. K., Yuryeva N. E. Stochastic swarm clusterization method in natural language data processing. // *Experimental Psychology (Russia)*. 2018. № 11(3), P. 5–18. DOI: 10.17759/exppsy.2018110301.
- [6] A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm / C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, J. Liu // *Eurasip Journal on Wireless Communications and Networking*. 2021. № 2021 (1). DOI: 10.1186/s13638-021-01910-w.
- [7] An improved DBSCAN method for LiDAR data segmentation with automatic Eps estimation / C. Wang, M. Ji, J. Wang, W. Wen, T. Li, Y. Sun // *Sensors (Switzerland)*. 2019. № 19(1). DOI: 10.3390/s19010172.