

# Иерархический агломеративный алгоритм кластеризации для выделения регионов изображений

О.Л. Вовк

Кафедра Прикладной математики и информатики  
Донецкий национальный технический университет, Донецк, Украина  
vovkolga@ukrtop.com

## Аннотация

В работе рассматривается применение статистических методов кластеризации для распознавания однородных областей (регионов) цветных изображений. Предлагается новый статистический алгоритм кластеризации изображений, основанный на иерархической агломеративной технологии. Определяется визуальная эффективность качества кластеризации разработанного алгоритма.

*Ключевые слова:* кластеризация, изображение, битовая маска взаимосвязей и рангов, контекстный поиск.

## 1. ВВЕДЕНИЕ

Постоянное увеличение технических возможностей современных вычислительных средств делает возрастными проблемами хранения и скорости обработки цветных изображений. Однако, одновременно с этим, возрастает потребность исследователей и пользователей в эффективных средствах обработки изображений.

Одна из основных областей обработки изображений – распознавание кластеров изображений по цветовому подобию. Данная проблема решается исследователями различных научных сфер [1, 2], к примеру, при контекстном поиске в электронных базах данных и при содержательной классификации изображений (медицинская диагностика, удаленное наблюдение, анализ документов). Однако, в зависимости от сферы применения и характера анализируемых изображений выделяемые кластеры могут представлять как отдельные объекты изображений (сегментирование изображений [3, 4]), так и значимые цвета изображений, отражающие их сущность (квантование цветов изображений [5]).

Цель предлагаемой работы – рассмотреть статистические методы кластеризации для выделения регионов изображений (без учета их местоположения на изображении); на основании статистических технологий разработать эффективный алгоритм выделения значимых областей (регионов, кластеров). В качестве основного применения рассматриваемого в работе алгоритма предлагается контекстный поиск изображений в базах данных [6].

## 2. ПОСТАНОВКА ЗАДАЧИ КЛАСТЕРИЗАЦИИ ИЗОБРАЖЕНИЙ

Кластеризация [7] – общее название множества вычислительных процедур, используемых при создании классификации объектов. В результате кластеризации образуются “кластеры” (регионы, области) – группы похожих по различным характеристикам объектов.

Исходные данные для процедуры кластеризации [7] – набор объектов, каждый из которых задается вектором своих характеристик. В ходе процедуры кластеризации происходит объединение “подобных” объектов в отдельные классы. Результатом кластеризации является набор классов, содержащих однородные объекты [6-8].

В случае решения задачи кластеризации изображения задан набор пикселей, каждый из которых определяется тремя цветовыми компонентами в одном из цветовых пространств. С помощью процедуры кластеризации осуществляется выделение групп пикселей, имеющих наиболее “близкие” цветовые компоненты.

Пример выделения регионов (кластеров) изображения приведен на рис. 1.

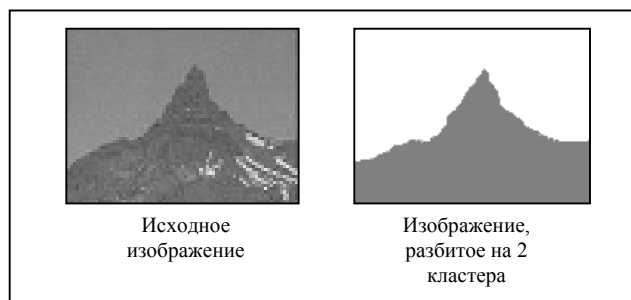


Рисунок 1: Пример кластеризации изображения

## 3. ОСНОВНЫЕ МЕТОДЫ СТАТИСТИЧЕСКОЙ КЛАСТЕРИЗАЦИИ

В [2, 4] выделяются две основные технологии статистической кластеризации: иерархическая и итеративная (разбиений).

Иерархические методы представляют собой процедуры создания последовательности вложенных разбиений, исходя из данных матрицы близости [2, 4].

По способу формирования кластеров иерархические методы подразделяются на методы одиночной связи и методы полной связи [9]. При одиночной связи – в один кластер объединяются объекты с минимальным расстоянием, а при полной связи – в разные кластеры разносятся объекты с максимальным расстоянием.

По способу инициализации первоначального разбиения на кластеры для иерархических технологий принято выделять две основные стратегии [2, 7]: агломеративную (при инициализации каждый объект – отдельный кластер) и дивизивную (изначально все объекты принадлежат одному кластеру).

**Таблица 1**

Младшие биты маски, характеризующие взаимосвязь цветковых компонентов

Компоненты	Связь цветковых характеристик	Маска
R и G	R>G	0 0 0 0 0 0 0 0 1
	R=G	0 0 0 0 0 0 0 1 0
	R<G	0 0 0 0 0 0 1 0 0
R и B	R>B	0 0 0 0 0 1 0 0 0
	R=B	0 0 0 0 1 0 0 0 0
	R<B	0 0 0 1 0 0 0 0 0
G и B	G>B	0 0 1 0 0 0 0 0 0
	G=B	0 1 0 0 0 0 0 0 0
	G<B	1 0 0 0 0 0 0 0 0

**Таблица 2**

Старшие биты маски, характеризующие уровни цветковых компонентов

Компонент	Предел изменения компонента	Ранг	Маска
B	$[x_l \dots GL]$	низкий	0 0 0 0 0 0 0 0 1
	$(GL \dots GH]$	средний	0 0 0 0 0 0 0 1 0
	$(GH \dots x_h]$	высокий	0 0 0 0 0 0 1 0 0
G	$[x_l \dots GL]$	низкий	0 0 0 0 0 1 0 0 0
	$(GL \dots GH]$	средний	0 0 0 0 1 0 0 0 0
	$(GH \dots x_h]$	высокий	0 0 0 1 0 0 0 0 0
R	$[x_l \dots GL]$	низкий	0 0 1 0 0 0 0 0 0
	$(GL \dots GH]$	средний	0 1 0 0 0 0 0 0 0
	$(GH \dots x_h]$	высокий	1 0 0 0 0 0 0 0 0

**Таблица 3**

Примеры битовых масок

Числовые значения центров кластеров			Маска
R	G	B	Биты
			17 ... 0
0.5	0.5	0.5	0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
0.7	0.1	0.3	1 1 0 0 0 1 0 1 1 1 0 0 0 0 1 0 0 1
0.2	0.8	1	0 0 1 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0
0.3	0.35	0.9	0 1 1 0 1 1 1 0 0 1 0 0 1 0 0 1 1 0
0.35	0.6	0.3	0 1 1 1 1 0 0 1 1 0 0 1 0 1 1 1 0 0

Разработанный иерархический алгоритм предназначен для кластеризации изображений в трехмерном цветовом пространстве RGB (т.е. для каждой точки изображения вектор характеристик содержит свойства красного, зеленого и синего компонентов) и состоит из трех основных этапов:

1. Этап полной связи. Предполагается, что изначально каждый пиксель изображения представляет собой отдельный кластер; на этом этапе происходит разнесение точек

В результате работы иерархических агломеративных методов, все объекты классификации будут принадлежать одному кластеру. Поэтому, для получения значимых разбиений, необходимо рассматривать различные “срезы” построенной иерархии [7, 9].

Основная идея итеративных методов (методов разбиения) – нахождение единственного разделения шаблонов (объектов) по кластерам путем отнесения каждого из объектов к кластеру, расстояние до центра которого минимально [2, 9].

Наиболее используемым методом данной группы является метод k-means кластеризации [1, 4, 9, 10], сравнение с которым проводится в данной работе.

#### 4. НОВЫЙ ИЕРАРХИЧЕСКИЙ АЛГОРИТМ ДЛЯ КЛАСТЕРИЗАЦИИ ИЗОБРАЖЕНИЙ

В основе предлагаемого алгоритма – битовая маска связей и рангов цветковых компонентов центров кластеров. Так как любое цветковое пространство трехмерно [11], то разработанная маска отражает взаимосвязи и ранги трех цветковых характеристик. Она содержит 18 бит, причем 9 младших бит характеризуют взаимосвязи цветковых составляющих, а старшие 9 бит отражают уровни (ранги) каждого из анализируемых компонентов в отдельности.

При формировании младших бит маски, каждая пара компонентов анализируется на наличие связей типа: меньше, больше и равно.

Старшие биты маски описывают уровни каждой из трех характеристик; вводится три основных уровня – низкий, средний и высокий. Для этого весь интервал изменения каждой из цветковых составляющих  $[x_l, x_h]$  делится на три равных промежутка –  $[x_l, GL]$ ,  $(GL, GH]$ ,  $(GH, x_h]$ , соответствующих приведенным уровням.

В таблицах 1, 2 приведены правила формирования младших и старших бит маски для пространства цветов RGB, т.е.  $x_r=0$  и  $x_g=1$  (минимальное и максимальное возможные значения цветковых компонентов в пространстве RGB, а не минимальное и максимальное значения цветковых характеристик анализируемого изображения).

Кроме того, предлагается “размывать” границы, как отношений, так и уровней, с помощью введения некоторой погрешности маски –  $\epsilon$ .

Т.е. в отношении между компонентами  $X_1$  и  $X_2$  присутствует связь, к примеру “>”, в случае, если удовлетворяется условие  $X_1 > X_2$ ; кроме того, предполагается, что если выполняется условие  $X_1 < (X_2 + \epsilon)$ , то между  $X_1$  и  $X_2$  существует и связь “=”. Аналогичными являются рассуждения и для отношения “<”.

Приведем анализ рангов цветковых характеристик. К примеру, пусть  $X_l = x_l$ , тогда:

- если  $x_l < (GL + \epsilon)$ , то  $X_l$  имеет низкий уровень;
- если  $x_l > (GH - \epsilon)$ , то  $X_l$  имеет высокий уровень;
- если  $(x_l > (GL - \epsilon) \text{ или } x_l < (GH + \epsilon))$ , то  $X_l$  имеет средний уровень.

Таблица 3 содержит примеры масок для конкретных числовых значений центров кластеров. Предполагается, что значения компонентов R, G и B принадлежат интервалу [0, 1],  $\epsilon=0.1$ ,  $GL=0.33$  и  $GH=0.67$ .

изображения в кластеры с одинаковыми битовыми масками. Т.е. для каждого пикселя изображения формируется маска взаимосвязей и рангов, затем пиксели с различными масками разносятся в разные кластеры.

2. Этап одиночной связи. Для полученных на предыдущем этапе кластеров строится матрица близости (расстояний) между кластерами. В качестве расстояния между кластерами используется среднее Евклидово расстояние между всеми парами точек, входящих в кластеры, расстояние между которыми рассчитывается. В полученной матрице происходит поиск наиболее “близких” кластеров (т.е. минимумов матрицы близости). Если расстояния между несколькими парами кластеров являются одинаковыми и минимальными, то, в первую очередь, в качестве “близких” кластеров, выбираются кластеры с минимальной площадью. Найденные “близкие” кластеры объединяются, образуя новые кластеры, для которых производится перерасчет центров. Из матрицы расстояний удаляются строки и столбцы, соответствующие объединенным кластерам, и добавляется строка и столбец, соответствующие полученному кластеру.

Стоит отметить, что первоначально, поиск наиболее “близких” кластеров предлагается производить для кластеров с незначительной площадью (автор работы незначительными определяет кластеры, площадь которых меньше 0.001% площади анализируемого изображения). Это позволяет минимизировать наличие эффектов “соли” и “перца” на кластеризованных изображениях.

Данный этап повторяется до тех пор, пока количество кластеров не достигнет количества шестнадцати (данный числовой параметр взят из работы [10] и проверен экспериментально).

3. Этап окончания. Данный этап разработан как критерий окончания кластеризации; он аналогичен предыдущему этапу с некоторым ограничением. Объединение кластеров с минимальным расстоянием матрицы близости производится только при выполнении условия “близости” масок; в противном случае процесс кластеризации заканчивается. Обозначим маски кластеров, претендующих на объединение –  $M_1$  и  $M_2$ . Тогда основные стадии анализа выполнения условия “близости” масок следующие:

- расчет результирующей маски объединения:  $M_r = M_1 \& M_2$ ;
- установка начального значения количества эквивалентных бит:  $K_b = 0$ ;
- анализ всех соответствующих триад масок  $M_r$ ,  $M_1$  и  $M_2$  в отдельности:

Если  $((M_r \& 7) \& \& !((M_1 \& 2 \parallel M_2 \& 2) \& \& (M_1 \& 4 \& \& M_2 \& 1) \parallel (M_2 \& 4 \& M_1 \& 1)))$

То  $K_b = K_b + 1$

- переход к следующей триаде масок:  $M_r = M_r \gg 3$ ,  $M_1 = M_1 \gg 3$  и  $M_2 = M_2 \gg 3$ ;

- возвращение к анализу следующей триады масок.
- анализ полученного числа эквивалентных битов:

Если  $(K_b > 5)$

То условие “близости” масок выполняется;

Иначе условие “близости” масок не выполняется.

Т.е. при анализе триад масок количество эквивалентных бит инкрементируется, если конъюнкция масок кластеров содержит хотя бы одну единицу ( $M_r \& 7$ ), причем пара триад рассматриваемых масок не должна быть двух следующих

типов  $(!(M_1 \& 2 \parallel M_2 \& 2) \& \& (M_1 \& 4 \& \& M_2 \& 1) \parallel (M_2 \& 4 \& M_1 \& 1))$ : 1) (1 1 0) и (0 1 1); 2) (0 1 1) и (1 1 0).

На рис.2 приведены примеры 24-битных изображений, размером 384x256 пикселей, различной энтропии. На рис.3 предлагаются визуальные результаты выделения значимых областей изображений с помощью авторского алгоритма, на рис. 4 – с помощью модификации k-means алгоритма [10].

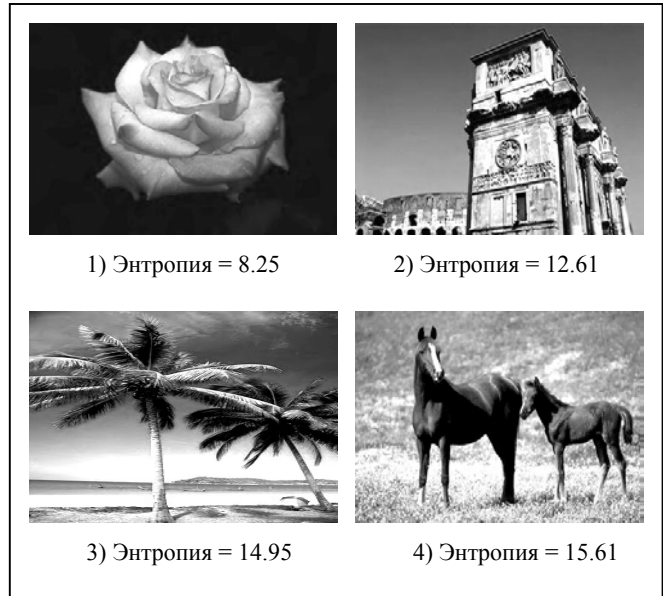


Рисунок 2: Примеры анализируемых изображений.



Рисунок 3: Примеры выделения регионов изображений с помощью предлагаемого иерархического алгоритма.



**Рисунок 4:** Примеры выделения регионов изображений с помощью k-means алгоритма.

Анализируя визуальные результаты кластеризации, можно сделать следующие выводы:

- рассматриваемый иерархический алгоритм менее чувствителен к последовательной смене оттенков одного цвета (изображения 2, 3);
- внутренние части кластеров, выделяемых авторским алгоритмом, более простые, т.е. с меньшим числом отверстий (изображения 1, 4);
- регионы, полученные при помощи предлагаемого алгоритма более однородные, т.е. содержат меньшее количество геометрически рассредоточенных групп небольшого числа пикселей (изображения 2, 4).

## 5. ЗАКЛЮЧЕНИЕ

Работа посвящена разработке нового статистического иерархического агломеративного алгоритма для выделения однородных областей цветных изображений. Предлагаемый алгоритм визуально сравнивается с наиболее распространенным алгоритмом кластеризации противоположной группы методов кластеризации – k-means алгоритмом, модификация которого для кластеризации изображений описана в [10]. Основными предпосылками визуальной эффективности разработанного алгоритма можно считать: введение битовой маски связей и рангов для предварительного разбиения и в качестве критерия окончания процесса кластеризации, “обязательное” объединение незначительных кластеров, использование в качестве расстояния между кластерами среднего Евклидова расстояния между всеми парами пикселей, содержащихся в сравниваемых кластерах.

В качестве основного направления дальнейших исследований выделяется анализ устойчивости разработанного алгоритма к вариации параметров освещенности и различным шумам.

## 6. ЛИТЕРАТУРА

- [1] Kanungo T., Mount D., Netanyahu N., Piatko C., Silverman R., Wu A. *An Efficient k-Means Clustering Algorithm: Analysis and Implementation* // IEEE Transactions on Pattern Analysis and Machine Intelligence. – July 2002. – vol. 24, №7. – P. 881-892.
- [2] Chen C.H., Pau L.F., Wang P.S.P. *The Handbook of Pattern Recognition and Computer Vision (2<sup>nd</sup> Edition)* – World Scientific Publishing Co., 1998. – 1004 p.
- [3] Deng Y., Manjunath B.S., Shin H. *Color Image Segmentation* // IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'99. – 1999. – vol. 2. – P. 446-451.
- [4] Zhong S., Ghosh J. *A Unified Framework for Model-based Clustering* // Journal of Machine Learning Research. – 2003. – №4. – P. 1001-1017.
- [5] Deng Y., Kenney S., Moore M.S., Manjunath B.S. *Peer group filtering and perceptual color image quantization* // Proc. IEEE International Symposium on Circuits and Systems VLSI. – Orlando, FL, June 1999. – vol. 4. – P. 21-24.
- [6] Вовк О.Л. *Поиск изображений в электронных коллекциях на основе регионального разбиения* // САИТ-2003. – Киев, 2003, с. 144-145.
- [7] Ким Д.О., Мьюллер Ч.У., Клекка У.Р. *Факторный, дискриминантный и кластерный анализ*. – М.: Финансы и статистика, 1989. – 215 с.
- [8] *Классификация и кластер* // Под ред. Райзина Д. В. – М.: Мир, 1980. – 393 с.
- [9] Jain A.K., Murty M.N., Flynn P.J. *Data Clustering: A Review* // ACM Computing Surveys. – 1999. – vol. 31, № 3. – P. 264-323.
- [10] Wang J. Z., Du Y. *Scalable Integrated Region-based Image Retrieval using IRM and Statistical Clustering* // Proc. ACM and IEEE Joint Conference on Digital Libraries. – Roanoke, VA, ACM, June 2001. – P. 268-277.
- [11] Прэтт У. *Цифровая обработка изображений*. – М.: Мир, 1982. – Кн. 1 – 312 с.

## Информация об авторе

Вовк Ольга Леонидовна – ассистент (аспирант) кафедры Прикладной математики и информатики Донецкого национального технического университета, Украина.

Научные интересы: контекстный поиск в электронных коллекциях изображений, выделение отдельных объектов изображений и анализ их характеристик (таких как: характеристики текстуры, формы, местоположения).

Контактная информация: 83000, Донецк, ул. Артема, д. 58, Донецкий национальный технический университет; телефон (рабочий): 91-08-56, e-mail: vovkolga@ukrtop.com.