УДК 681.3

# ОЦЕНКА ЭФФЕКТИВНОСТИ НОВОГО СТАТИСТИЧЕСКОГО ИЕРАРХИЧЕСКОГО АГЛОМЕРАТИВНОГО АЛГОРИТМА КЛАСТЕРИЗАЦИИ ДЛЯ РАСПОЗНАВАНИЯ РЕГИОНОВ ИЗОБРАЖЕНИЙ

**Е. А. БАШКОВ, О. Л. ВОВК**

Рассматривается новый иерархический агломеративный алгоритм выделения кластеров (регионов) изображений. Приводится анализ быстродействия предлагаемого алгоритма в сравнении с наиболее распространенным алгоритмом *k*-means кластеризации. Вводятся оценки качества кластеризации изображений.

## ВВЕДЕНИЕ

На современном уровне развития технических средств повышается интерес к нетрадиционным сферам внедрения компьютерных технологий. К их числу относятся и сферы использования методов распознавания объектов изображений. Задачи выделения таких объектов (задачи кластеризации) наиболее часто решаются при контекстном поиске в электронных базах данных и при содержательной классификации изображений (медицинская диагностика, удаленное наблюдение, анализ документов) [1, 2].

Цель данной статьи – оценить эффективность нового статистического алгоритма кластеризации путем введения различных оценок затрат процессорного времени и качества кластеризации.

В соответствии с поставленной целью в предлагаемой работе решаются следующие основные задачи:

- рассматривается общая постановка задачи кластеризации;
- предлагается новый алгоритм кластеризации, являющийся модификацией статистического иерархического агломеративного алгоритма распознавания однородных областей;
- вводится теоретическая оценка нового алгоритма по критерию затрат процессорного времени (рассматривается роль сегментации при кластеризации изображений);
- проводится экспериментальный анализ качества кластеризации авторского алгоритма по критериям быстродействия и качества.

## 1. ПОСТАНОВКА ЗАДАЧИ КЛАСТЕРИЗАЦИИ

Кластеризация [3] — общее название множества вычислительных процедур, используемых при создании классификации объектов, в результате которых образуются «кластеры» (регионы) — группы похожих по различным характеристикам объектов. Кластерный метод [3] — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о характеристиках объектов, и затем упорядочивающая объекты в сравнительно однородные группы.

Исходные данные для процедуры кластеризации [4]: набор объектов, каждый из которых задается вектором своих характеристик. Существует два основных формата представления входных данных процедуры кластеризации [2, 5] — матрица шаблонов и матрица близости.

В ходе процедуры кластеризации происходит объединение «подобных» объектов в отдельные классы. Результат кластеризации — набор классов, содержащих однородные объекты [3, 4].

В задаче выделения объектов изображения задан набор пикселей, каждый из которых определяется тремя цветовыми компонентами в одном из цветовых пространств. С помощью процедуры кластеризации выделяются группы пикселей, имеющие наиболее близкие цветовые компоненты, т.е. происходит распознавание изображений (рис. 1).

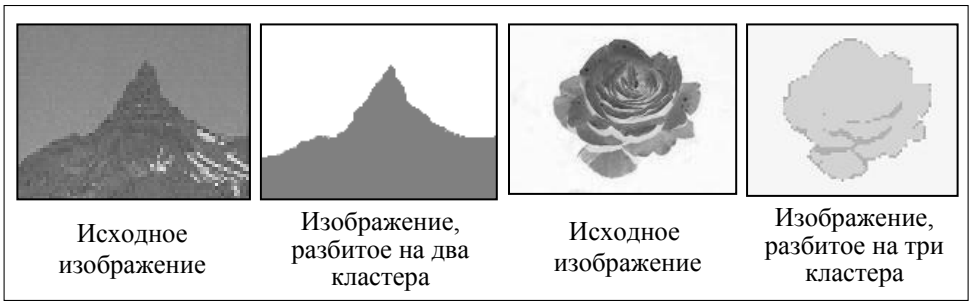



*Рис. 1*. Примеры кластеризации изображений

## 2. ОСНОВНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ КЛАСТЕРИЗАЦИИ

В соответствии с работами [3, 4] все разработанные кластерные методы статистики образуют шесть основных групп:

1. **Иерархические агломеративные методы**. В начале процедуры кластеризации каждый объект представляет собой отдельный кластер, затем происходит объединение наиболее «близких» кластеров. Процесс объединения продолжается до тех пор, пока не будет удовлетворено условие окончания кластеризации.

2. **Иерархические дивизимные методы** являются логической противоположностью методов предыдущей группы. При инициализации процесса кластеризации все объекты принадлежат одному классу, а затем этот всеобъемлющий кластер делится на последовательно уменьшающиеся части.

3. **Итеративные методы группировки.** Предварительно все данные разбиваются на некоторое количество кластеров, для которых вычисляются центры тяжести. Каждый объект помещается в кластер с ближайшим

центром тяжести. Вычисляются новые центры тяжести. Эта процедура продолжается до тех пор, пока не перестанут меняться кластеры.

4. **Методы поиска модальных значений плотности.** Рассматривается кластер как область пространства с высокой плотностью точек по сравнению с окружающими областями. В ходе кластеризации происходит поиск скоплений данных, которые и представляют собой области высокой плотности.

5. **Факторные методы** основываются на формировании корреляционной матрицы сходств между объектами, по которой определяются факторы, и объекты распределяются по кластерам в зависимости от их факторных нагрузок.

6. **Методы сгущений**. Уникальность их состоит в том, что они позволяют создавать перекрывающиеся кластеры. Основываются на вычислении матрицы сходства между объектами и определении оптимального значения статистического критерия, называемого специалистами «функцией сцепления». Затем объекты перемещаются до тех пор, пока функция не достигнет оптимального значения.

Рассмотренные группы методов соответствуют разным подходам к созданию кластеров. Применение различных методов к одним и тем же данным может привести к существенно различным результатам [3]. Поэтому особое внимание необходимо уделить выбору метода кластеризации для каждой конкретной задачи разделения объектов на однородные группы.

При решении проблемы кластеризации изображений заданный набор точек рассматривается как многомерная совокупность признаков объектов, согласно которым требуется рассортировать эти объекты по классам. С учетом исходных данных не все приведенные выше группы методов применимы к корректному решению задачи выделения объектов изображений.

Например, методы поиска модальных значений трудно применимы к кластеризации изображений, так как они основаны на объединении кластеров по некоторому заданному правилу, которое необходимо изменять при анализе каждого нового изображения.

Факторные методы предполагают, что объекты распределяются по кластерам в зависимости от факторных нагрузок. В случае обработки изображений в качестве факторов выступают три цветовые компоненты. Согласно факторному методу по отдельным кластерам объекты распределяются с высокой нагрузкой по каждой из трех компонент. Однако алгоритмы этой группы не учитывают ситуацию, когда высокая нагрузка присутствует по нескольким компонентам (при анализе изображений — это обычный случай).

Основное достоинство методов сгущения — возможность создания перекрывающихся кластеров — является основным недостатком при кластеризации изображений. При выделении объектов изображений ставится требование однозначности, а, следовательно, один объект не может быть членом нескольких кластеров.

Таким образом, для выделения регионов изображений наиболее применимы иерархические и итеративные методы. Именно эти две группы методов рассматриваются как основные в работах [2, 5].

Иерархические методы представляют собой процедуры создания последовательности вложенных разбиений, исходя из данных матрицы близости [2]. Формально любой иерархический метод кластеризации состоит из следующих шагов [5]:

1. Расчет матрицы близости между всеми парами шаблонов. Изначально каждый объект — отдельный кластер.

2. Нахождение минимума в матрице близости и объединение кластеров с минимальным расстоянием. Обновление строк и столбцов матрицы, соответствующих объединенным кластерам.

3. Если все объекты принадлежат одному кластеру, то конец работы метода. Иначе на шаг 2.

В результате работы таких методов все объекты классификации будут принадлежать одному кластеру. Поэтому для получения значимых разбиений необходимо рассматривать различные срезы построенной иерархии [2].

По способу формирования кластеров иерархические методы подразделяются на методы одиночной и методы полной связи [5]. При одиночной связи в один кластер объединяются объекты с минимальным расстоянием, а при полной — в разные кластеры разносятся объекты с максимальным расстоянием.

Основная идея итеративных методов (методов разбиения) [5] — нахождение единственного разделения шаблонов по кластерам вместо иерархии, полученной согласно иерархическим технологиям. Реализация методов разбиения предполагает выполнение таких шагов [2, 5]:

1. Выбор начального распределения объектов по кластерам. Расчет центров тяжести полученных кластеров.

2. Перегруппировка объектов кластеров (отнесение каждого объекта к кластеру с минимальным расстоянием до центра).

Однако такая технология имеет существенные ограничения [5]: при разных стартовых условиях такие методы выдают различные результаты, нет методики выбора количества кластеров.

Наиболее используемый метод данной группы — метод *k*-means кластеризации [1, 5–7], который при решении задачи выделения объектов изображений состоит из следующих основных этапов:

1. Произвольное разбиение на некоторое заданное количество кластеров (обычно изначально количество кластеров равно двум).

2. Вычисление центров тяжести полученных кластеров.

3. Анализ каждого объекта каждого кластера. Перераспределение в кластер с минимальным расстоянием до центра.

4. После перераспределения — пересчет центров тяжести кластеров.

5. Этапы 3, 4 повторяются до тех пор, пока все точки не будут находиться в «ближайшем» кластере.

6. Если не удовлетворен критерий окончания кластеризации, то увеличивается количество кластеров, повторяются шаги 1 – 5.

Необходимо отметить, что в работах [6, 7] показано, что для выделения наиболее информативных областей изображений при их кластеризации количество кластеров необходимо ограничивать шестнадцатью.

## 3. НОВЫЙ СТАТИСТИЧЕСКИЙ АГЛОМЕРАТИВНЫЙ АЛГОРИТМ (НСАА) ДЛЯ КЛАСТЕРИЗАЦИИ ИЗОБРАЖЕНИЙ

### 3.1. Описание НСАА

В основе предлагаемого алгоритма — битовая маска связей и рангов цветовых компонентов центров кластеров. Так как любое цветовое пространство

трехмерно [8], то разработанная маска отражает взаимосвязи и ранги трех цветовых характеристик. Она содержит 18 бит, причем 9 младших бит характеризуют взаимосвязи цветовых составляющих, а старшие 9 —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 цветовых составляющих $[x_l, x_h]$ делится на три равных промежутка $[x_l, GL]$, $[GL, GH]$, $(GH, x_h]$, соответствующих приведенным уровням.

В табл. 1 приведены правила формирования младших и старших бит маски для пространства цветов RGB.

Кроме того, предлагается «размывать» границы как отношений, так и уровней с помощью введения некоторой погрешности маски $\varepsilon$. Т.е. в отношении между компонентами $X_1$ и $X_2$ присутствует связь, например, «$>$», если удовлетворяется условие $X_1 > X_2$. Кроме того, предполагается, что если выполняется условие $X_1 < (X_2 + \varepsilon)$, то между $X_1$ и $X_2$ существует и связь «$=$». Аналогичны рассуждения и для отношения «$<$».

**Пример** анализа рангов цветовых характеристик.

Пусть $X_1 = x_1$, тогда:

- если $x_1 \le (GL + \varepsilon)$, то $X_1$ имеет низкий уровень;
- если $x_1 \ge (GH - \varepsilon)$, то $X_1$ имеет высокий уровень;
- если $x_1 > (GL - \varepsilon)$ или $x_1 < (GL + \varepsilon)$, то $X_1$ имеет средний уровень.

В табл. 2 приведены примеры масок для конкретных числовых значений. Предполагается, что значения компонентов $R$, $G$, $B$ принадлежат интервалу $[0,1]$; $\varepsilon = 0{,}1$; $GL = 0{,}33$; $GH = 0{,}67$.

**Таблица 1.** Маска связей и рангов компонентов пространства *RGB*

| Старшие биты | | | | Младшие биты | | |
|---|---|---|---|---|---|---|
| Компонент | Предел изменения компонента | Ранг | Маска | Компоненты | Связь цветовых характеристик | Маска |
| *B* | $[x_l ... GL]$ | низкий | 0 0 0 0 0 0 0 0 1 | *R* и *G* | $R > G$ | 0 0 0 0 0 0 0 0 1 |
| | $(GL ... GH]$ | средний | 0 0 0 0 0 0 0 1 0 | | $R = G$ | 0 0 0 0 0 0 0 1 0 |
| | $(GH ... x_h]$ | высокий | 0 0 0 0 0 0 1 0 0 | | $R < G$ | 0 0 0 0 0 0 1 0 0 |
| *G* | $[x_l ... GL]$ | низкий | 0 0 0 0 0 1 0 0 0 | *R* и *B* | $R > B$ | 0 0 0 0 0 1 0 0 0 |
| | $(GL ... GH]$ | средний | 0 0 0 0 1 0 0 0 0 | | $R = B$ | 0 0 0 0 1 0 0 0 0 |
| | $(GH ... x_h]$ | высокий | 0 0 0 1 0 0 0 0 0 | | $R < B$ | 0 0 0 1 0 0 0 0 0 |
| *R* | $[x_l ... GL]$ | низкий | 0 0 1 0 0 0 0 0 0 | *G* и *B* | $G > B$ | 0 0 1 0 0 0 0 0 0 |
| | $(GL ... GH]$ | средний | 0 1 0 0 0 0 0 0 0 | | $G = B$ | 0 1 0 0 0 0 0 0 0 |
| | $(GH ... x_h]$ | высокий | 1 0 0 0 0 0 0 0 0 | | $G < B$ | 1 0 0 0 0 0 0 0 0 |

**Таблица 2.** Примеры битовых масок

| Числовые значения центров кластеров | | | Маска | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *R* | *G* | *B* | Биты | | | | | | | | | | | | | | | | | |
| | | | 17 | 16 | 15 | 14 | 13 | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 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 состоит из трех основных этапов.

1. **Полная связь.** Предполагается, что изначально каждый пиксель изображения представляет собой отдельный кластер. Происходит разнесение точек изображения в кластеры с одинаковыми битовыми масками. Т.е. для каждого пикселя изображения формируется маска взаимосвязей и рангов, затем пиксели с одинаковыми масками разносятся в разные кластеры.

2. **Одиночная связь.** Для полученных на предыдущем этапе кластеров строится матрица близости (расстояний) между кластерами. В качестве расстояния используется среднее евклидово расстояние между всеми парами точек, входящих в кластеры с рассчитываемыми между ними расстояниями. В полученной матрице происходит поиск наиболее «близких» кластеров (т.е. кластеров с минимальным расстоянием). Они объединяются, образуя новые кластеры, для которых производится перерасчет центров. Причем $i$-й и $j$-й кластеры считаются «близкими» только в том случае, если расстояние от $i$-го кластера к $j$-му минимально в $i$-й строке матрицы близости, и расстояние от $j$-го кластера к $i$-му минимально в $j$-й строке матрицы близости. Из матрицы расстояний удаляются строки и столбцы, соответствующие объединенным кластерам, и добавляется строка и столбец, соответствующие полученному кластеру. Данный этап повторяется до тех пор, пока количество кластеров не достигнет шестнадцати (данный числовой параметр взят из работ [6, 7] и проверен экспериментально).

3. **Окончание**. Данный этап разработан как критерий окончания кластеризации. Он аналогичен предыдущему 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 сформулируем так:

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

если $\big((M_r \,\&\, 7)\&\,\&!\big((M_1 \,\&\, 2 \| M_2 \,\&\, 2)\&\,\&\big((M_1 \,\&\, 4\&\,\&M_2 1)\|(M_2 \,\&\, 4\,\&\,\&\, M_1 \,\&\, 1)\big)\big)\big)$,

то $K_b = K_b + 1$;

- переход к следующей триаде масок: $M_r = M_r >> 3$, $M_1 = M_1 >> 3$ и $M_2 = M_2 >> 3$;
- возвращение к анализу следующей триады масок;
- анализ полученного числа эквивалентных битов:

если ($K_b \geq 5$), то условие «близости» масок выполняется, иначе — 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 $(!((M_1 \,\&\, 2 \| M_2 \,\&\, 2) \& \& ((M_1 \,\&\, 4 \& \& M_2 1) \| (M_2 \,\&\, 4 \& \& M_1 \,\&\, 1))))$:

1) (1 1 0) и (0 1 1);
2) (0 1 1) и (1 1 0).

### 3.2. Визуальные результаты кластеризации

На рис.2 показаны результаты выделения значимых областей изображений с помощью авторского алгоритма. В качестве тестируемых изображений выбраны 24-битовые изображения размером 512x512 пикселей: baboon.tiff — энтропия 17,74; lena.tiff — энтропия 16,84.

Исходное изображение | Изображение, разбитое на четыре кластера, процессорное время 6890 мс | Исходное изображение | Изображение, разбитое на три кластера, процессорное время 5625 мс

*Рис. 2*. Примеры распознавания изображений с помощью алгоритма HCAA

На рис.3 приведены результаты кластеризации тех же изображений с помощью противоположной (итеративной) технологии. Как пример итеративной

Исходное изображение | Изображение, разбитое на пять кластеров, процессорное время 997614 мс | Исходное изображение | Изображение, разбитое на три кластера, процессорное время 825652 мс

*Рис. 3*. Примеры распознавания изображений с помощью алгоритма *k*-means

группы методов кластеризации выбран алгоритм *k*-means, модификация которого для выделения объектов изображений подробно описана в работах [6,7].

### 3.3. Анализ затрат оперативной памяти

При обработке изображений начальное количество кластеров будет $k = hw$, где $h$ — высота изображения; $w$ — ширина изображения в пикселях.

Тогда для хранения матрицы близости с элементами заданного типа требуемый объем оперативной памяти в байтах можно вычислить по формуле

$$V_{\text{ОП}} = k^2 b, \tag{1}$$

где $b$ — количество байт, необходимое для хранения одного элемента заданного типа.

Самый очевидный способ минимизации этого объема — хранение не всей матрицы, а нижней (верхней) треугольной ее части без диагонали, что возможно в силу симметричности матрицы. Необходимое количество байт оперативной памяти можно представить как

$$V_{\text{ОП}} = \frac{k^2 b}{2} - \frac{kb}{2}. \tag{2}$$

### 3.4. Анализ скорости кластеризации

Рассмотрим теоретическое обоснование эффективности разработанного метода по сравнению с алгоритмом *k*-means кластеризации [6, 7].

Пусть для заданного изображения размером $w \times h$ оптимальное количество кластеров согласно критерию окончания кластеризации — $n$. Оценим количество итераций, необходимое каждому из сравниваемых алгоритмов для достижения этого количества групп однородных элементов.

Рассмотрим применение алгоритма $k$-means [6, 7] для распознавания объектов изображений. Задачу вычисления максимального количества итераций, которое будет затрачено на выделение $n$ объектов изображения, можно свести к комбинаторной задаче размещений. Следовательно, максимальное количество итераций будет

$$I_{k-\text{means}} = \sum_{i=2}^{n} \frac{(wh)!}{(wh-i)!}. \tag{3}$$

Используя обозначения, введенные выше, максимальное число итераций для разработанного НСАА можно вычислить по формуле

$$I_{\text{НСАА}} = wh - n. \tag{4}$$

Необходимо отметить, что в формуле (4) произведена теоретическая оценка количества итераций предлагаемого алгоритма, полученная из предположения, что на первом этапе иерархического алгоритма (этапе полной связи) маски всех кластеров были различны, т.е. не было первоначального разнесения точек в кластеры с одинаковыми масками и, следовательно, количество кластеров изначально не уменьшалось. Такая стратегия оценки выбрана для предупреждения спора о природе первого этапа алгоритма.

Предполагается, что этап полной связи алгоритма может быть отнесен как к предварительной сегментации (выделению однородных областей — уменьшению количества первоначальных цветов изображений [6, 7, 9]), так и к предлагаемому алгоритму распознавания изображений непосредственно. При разработке алгоритма авторы склонялись ко второму варианту, так как на этапе полной связи использовали одну из особенностей иерархической стратегии кластеризации.

Практическое тестирование НСАА, проводимое на ПЭВМ Intel Celeron 2700 на коллекции из 1000 картинок, показало превосходство разработанного алгоритма над наиболее распространенным алгоритмом кластеризации *k*-means по критерию затрат процессорного времени (рис. 4, кривая алгоритма НСАА практически совпадает с осью абсцисс). Так как значения затрат процессорного времени несоизмеримы для сравниваемых алгоритмов (при таких шкалах осей показатели быстродействия НСАА не видны), то рекомендуется анализировать быстродействие НСАА по рис. 5. Анализируемые 24-битовые картинки были упорядочены по возрастанию энтропии (меры информативности изображений [8]). Показатели энтропии для тестируемых изображений приведены на оси абсцисс. Неравномерность нанесения значений по оси абсцисс связана с неравным количеством изображений различных показателей энтропии в экспериментальной базе данных (большинство изображений имеют энтропию в пределах от 14 до 15). Количество пикселей тестируемых изображений от 98304 ($384 \times 256$) до 262144 ($512 \times 512$) пикселей.

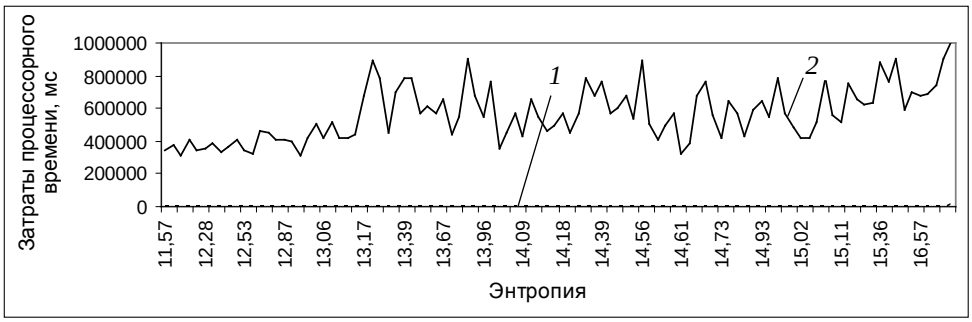



*Рис. 4.* Сравнение показателей затрат процессорного времени НСАА (*1*) и *k*-means (*2*) алгоритмов

Для чистоты эксперимента сравнение затрат процессорного времени проведено также для алгоритма *k*-means с изначальным разбиением по технологии этапа полной связи разработанного иерархического алгоритма НСАА с помощью битовой маски (результаты тестирования приведены на рис. 5).

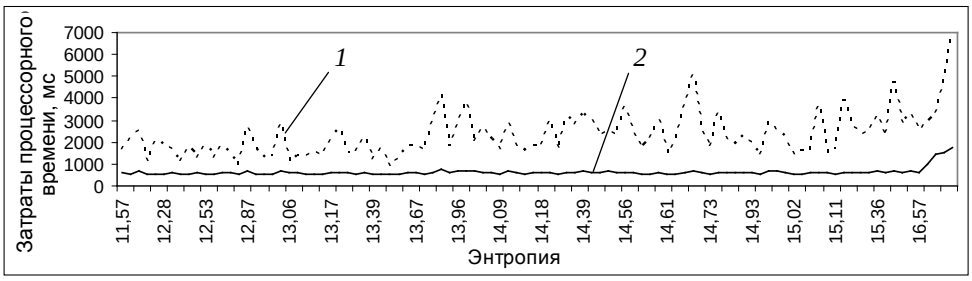



Рис. 5. Сравнение показателей затрат процессорного времени НСАА (*1*) и усовершенствованного *k*-means (*2*) алгоритмов

Анализ полученных результатов экспериментов показывает значимость введенной авторами битовой маски связей и рангов, при использовании которой алгоритм *k*-means получает ускорение в несколько сотен раз и превосходит даже разработанный алгоритм (HCAA). Введение битовой маски как этапа алгоритма *k*-means визуально практически не изменяет качество кластеризации. То же можно заметить и для критериев качества, описанных ниже.

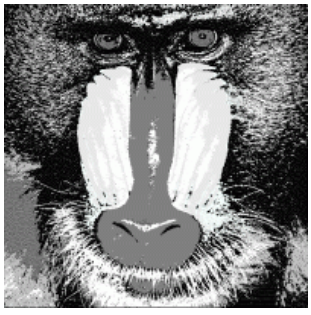
Отметим, что всевозможные «скачки» на рис. 4 и 5 связаны с различным числом выделяемых кластеров (ведь увеличение количества кластеров на единицу приводит к увеличению количества итераций в алгоритме *k*-means в среднем в два раза) и различными размерами анализируемых изображений.

### 3.5. Роль предварительной сегментации изображений

Под предварительной сегментацией будем понимать дискретизацию (квантование) цветов изображения для уменьшения исходного числа кластеризируемых цветов (кластеров) и, соответственно, ускорения процедуры выделения значимых объектов изображений. Как отмечалось выше, предварительной сегментацией условно можно считать и первый этап HCAA.

Реальные изображения наряду с полезной информацией содержат различные помехи. Источниками помех являются собственные шумы фотоприемных устройств, зернистость фотоматериалов, шумы каналов связи [8]. Предварительная сегментация уместна только в том случае, если ее введение не ухудшает качество кластеризации (т.е. помехи изображений не подчеркиваются, а сглаживаются).

Большинство методов сегментации основано на специфических свойствах различных цветовых пространств (*LUV*, *YUV*, *LAB*, *HSL*) [2, 10, 11]. Такие методы выделяют наиболее значимые цвета изображений, которые для низкоэнтропийных изображений и являются результирующими значимыми кластерами, а для изображений с высокой энтропией предполагают реализацию процедуры кластеризации со значительно уменьшенным исходным набором кластеров. Например, в основе *YUV* (сегментации [11]) лежит последовательное применение к палитре исходных цветов изображения 2*D* алгоритма *k*-means для характеристик *U* и *V* цветового пространства и 1*D* алгоритма *k*-means для характеристики *Y* (рис. 6).



Изображение, разбитое на пять кластеров, процессорное время 178307 мс

Изображение, разбитое на три кластера, процессорное время 126132 мс

*Рис. 6.* Пример предварительной YUV-сегментации

При простейшей предварительной сегментации [6, 7] исходное изображение разбивается на блоки заданного размера (в предлагаемых работах — блоки $4 \times 4$ пикселя). Цветовые характеристики всех пикселей полученных блоков принимаются равными средним цветовым характеристикам блоков (рис. 7).

Изображение, разбитое на четыре кластера, процессорное время 5500 мс

Изображение, разбитое на три кластера, процессорное время 1813 мс

*Рис. 7.* Пример кластеризации с предварительным разбиением на блоки $4 \times 4$

Анализируя результаты примеров кластеризации с предварительной сегментацией, необходимо заметить, что наряду со значительным ускорением процесса кластеризации, получаемом при сегментации на блоки, эта методика проигрывает по критерию качества. Таким образом, выбор технологии сегментации необходимо осуществлять, исходя из требований к качеству и доступных вычислительных мощностей. В нашей работе предлагаемый алгоритм не дополняется ни одним из видов сегментирования.

## 4. КРИТЕРИИ ОЦЕНКИ КАЧЕСТВА КЛАСТЕРИЗАЦИИ ИЗОБРАЖЕНИЙ

### 4.1 Функционал качества разбиения Колмогорова

Предлагается подход оценки качества разбиения изображений на кластеры, основанный на схеме, предложенной А.Н. Колмогоровым и описанной в работе [12]. Эта схема основана на понятии меры концентрации $Z(S)$ точек, соответствующей разбиению $S$, и средней меры внутриклассового рассеяния $I(S)$, характеризующей то же разбиение $S$.

Мера концентрации точек кластеров в данной схеме

$$Z(S) = \frac{1}{n} \sum_{i=1}^{k} n_i^2 , \qquad (5)$$

где $k$ — число различных кластеров в разбиении $S$; $n$ — общее число элементов во всех кластерах; $n_i$ — число элементов в кластере $S_i$.

Заметим, что предложенная мера концентрации имеет минимальное значение, равное $1/n$, при разбиении исследуемого множества на $n$ одноточечных кластеров, и максимальное значение, равное 1, при объединении всех объектов в один кластер.

Средняя мера внутриклассового рассеяния определяется

$$I(S) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\upsilon(X_i)}\sum_{X_l \in S(X_i)} d(X_i, X_l), \qquad (6)$$

где $\upsilon(X_i)$ — число элементов в кластере, содержащем точку $X_i$; $d(X_i, X_l)$ — расстояние между точками $X_i$ и $X_l$.

Определим функционал качества Колмогорова для распознавания образов изображений: качество разбиения изображения на кластеры выше у того из сравниваемых алгоритмов, для которого выражение $K(S) = 1/(1/Z(S) + I(S)) = Z(S)/(1 + I(S))$ имеет максимальное значение, т.е. при максимальном количестве точек кластера разброс значений цветовых характеристик является минимальным.

### 4.2. «Удаленность» кластеров друг от друга

При удовлетворении требований полноты и точности кластеризации результирующие кластеры должны не только иметь минимальный разброс цветовых характеристик (согласно критерию, описанному выше), но и быть максимально «удаленными» по отношению друг к другу.

В соответствии с [13] «удаленность» кластеров друг от друга можно измерять величиной

$$f = \frac{1}{k-1}\sum_{i=1}^{k-1} d_i, \qquad (7)$$

где $k$ — количество кластеров, полученное в результате разбиения; $d_i$ — расстояние между центрами кластеров.

Согласно этому критерию более эффективным является тот алгоритм кластеризации, для которого числовое значение $f$ будет наибольшим.

### 4.3. Учет значимости неиспользуемых характеристик

Критерием является результат проведения теста значимости, с помощью которого анализируются кластеры по признакам, не применявшимся при получении кластерного решения [3].

Для решения проблемы кластеризации изображений этот критерий предлагается переформулировать следующим образом.

Определить различия $R$ в кластерных решениях, найденные одним и тем же алгоритмом кластеризации, но с использованием различных цветовых характеристик. Т.е. необходимо сравнить решения, полученные каждым из исследуемых методов для различных цветовых пространств. Тот алгоритм, для которого данное различие минимально, и является более устойчивым к смене цветовых координат. Таким образом, показатель качества согласно этому критерию выше у алгоритмов с максимальным значением величины $1/R$.

В качестве анализируемых цветовых пространств выбраны пространства *RGB* и *HSL* [8, 14].

### 4.4. Сравнительный анализ качества алгоритмов кластеризации

Исследования качества алгоритмов кластеризации *k*-means и HCAA проводилось на базе данных картинок, описанных в подпункте 3.4. Результаты тестирования по критериям, рассмотренным выше, приведены на рис. 8–10.

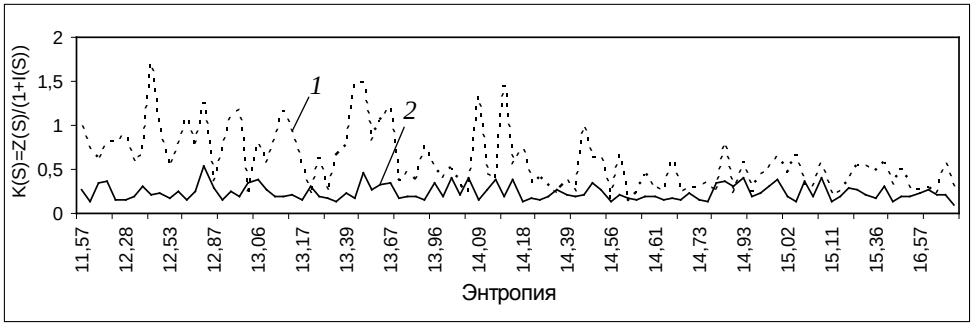



*Рис. 8*. Показатели функционалов качества разбиения Колмогорова: *1* — HCAA; *2* — *k*-means


*Рис. 9*. Показатели «удаленности» кластеров друг от друга: *1* — HCAA; *2* — *k*-means

Полученные практические результаты тестирования рассматриваемых алгоритмов показывают, что в абсолютном большинстве случаев лучшими являются показатели эффективности HCAA. Основными предпосылками получения таких числовых значений предлагаемых критериев можно считать:

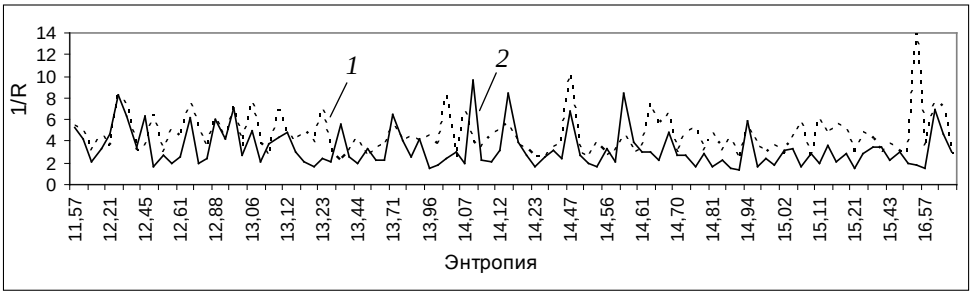



*Рис. 10*. Показатели критерия учета значимости неиспользуемых характеристик: *1* — HCAA; *2* — *k*-means

- отсутствие в алгоритме *k*-means учета вклада каждой из анализируемых характеристик (в методе HCAA эта задача решается введением маски связей и рангов цветовых компонентов);
- окончание процесса кластеризации в методе *k*-means по критериям, не связанным с данными конкретного рассматриваемого изображения в отличие от разработанного алгоритма [6, 7].

## ВЫВОДЫ

Предложен эффективный алгоритм кластеризации изображений НСАА. В основе нового метода кластеризации — статистический иерархический агломеративный метод. Для выделения объектов изображений проведена существенная модификация алгоритма, базирующаяся на введении маски взаимосвязей и рангов отдельных цветовых компонентов изображений, необходимой для учета значения каждого из этих компонентов.

В работе проведено многостороннее тестирование НСАА. Среди критериев тестирования: быстродействие, критерий функционалов качества разбиения Колмогорова, «удаленности» кластеров друг от друга и учета значимости неиспользуемых цветовых характеристик. В ходе практического тестирования НСАА и алгоритмов *k*-means более высокие показатели качества и быстродействия были получены для разработанного алгоритма. Для подтверждения практических результатов приведено теоретическое обоснование эффективности НСАА по предлагаемым критериям.

## ЛИТЕРАТУРА


1. *An Efficient* k-Means Clustering Algorithm: Analysis and Implementation / T. Kanungo, D. Mount, N. Netanyahu et al. // IEEE Transactions on Pattern Analysis and Machine Intelligence. — July 2002. — **24**, № 7. — P. 881–892.
2. *Chen C.H., Pau L.F., Wang P.S.P.* The Handbook of Pattern Recognition and Computer Vision (2nd Edition). — World Scientific Publishing Co. — 1998. — 1004 p.
3. *Ким Д.О., Мьюллер Ч.У., Клекка У.Р.* Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989. — 215 с.
4. *Классификация* и кластер / Под ред. Д. В. Райзина — М.: Мир, 1980. — 393 с.
5. *Jain A.K., Murty M.N., Flynn P.J.* Data Clustering: A Review // ACM Computing Surveys. — 1999. — **31**, № 3. — P. 264–323.
6. *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.
7. *Wang J.Z., Li J., Wiederhold G.* SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2001. — **23**, № 9. — P. 947–963.
8. *Прэтт У.* Цифровая обработка изображений. – Кн. 1. — М.: Мир, 1982. — 312 с.
9. *Путятин Е.П.* Нормализация и распознавание изображений. http: // sumschool.sumdu.edu.ua / i s-02 / rus / lectures / pytyatin / pytyatin.htm.
10. *Comaniciu D., Merr P.* Robust Analysis of Feature Spaces: Color Image Segmentation // Proc. Of CVPR'97. — P. 750–755.
11. *Lucchese L., Mitra S.K.* Unsupervised Segmentation of Color Images Based on *k*-means Clustering in the Chromaticity Plane // Proc. of IEEE Workshop on Content-based Access of Images and Video Libraries (CBAIVL'99). — Fort Collins, CO. — 1999. — P. 74–78.
12. *Айвазян С. А., Мхитарян В.С.* Прикладная статистика и основы эконометрики: Учебник для вузов. — М.: ЮНИТИ, 1998. — 1022 с.
13. *Загоруйко Н.Г.* Методы распознавания и их применение. — М.: Сов. радио, 1972. — 206 с.
14. *Руководство* по работе с цветом компании X-Rite. http://www.Realcolor.ru.


*Поступила 28.08.2003*