

Алгоритмы обработки изображений для классификации состояний биологических систем

В.В. Мажуга, М.В. Хачумов

Аннотация. Работа посвящена диагностике биологических систем на примере распознавания мочекаменной болезни человека на основе анализа образцов кристаллизованных капель биологических жидкостей, представленных в виде цветных и полутоновых изображений.

Ключевые слова: алгоритм, обработка, биологическая жидкость, система, изображение, признак, обработка, распознавание.

Введение

Ежедневно базы данных больниц и клиник пополняются терабайтами новых снимков – результатов обследований пациентов. В связи с этим возникает потребность в создании механизмов автоматического интеллектуального анализа изображений, способных оказать существенную поддержку работе врача. В последнее время получили развитие кристаллографические методы, применяемые в медицине и биологии. Они направлены на изучение структуры кристаллов, полученных на основе капель биологических жидкостей, в частности, мочи, крови, слюны – фаций [1]. Кристаллы изучают под микроскопом для выявления особенностей, что требует существенных затрат времени врача. Ставится задача автоматической обработки снимков фаций мочи для выявления наличия мочекаменной болезни.

1. Формализация представления изображения

Цифровое изображение хранится в виде двумерного массива, каждый элемент (m, n)

которого представляет собой пиксель с интенсивностью $I(m, n)$, изменяющейся в диапазоне от 0 до $L-1$. Величина L обычно является степенью двойки (например, 64, 256) и называется глубиной изображения [2]. Будем рассматривать функцию яркости изображения как стационарный случайный процесс [3]. В этом случае искомыми признаками для каждого снимка фации будут служить числовые характеристики случайного процесса. Для анализа часто используются гистограммы распределения значений яркости на изображениях, начальные и центральные моменты. Построим нормализованную гистограмму для каждого исходного изображения $p(z_i) = \frac{n_i}{n}$, где n_i – число пикселей с яркостью z_i ($i = 0, \dots, L-1$), n – общее число пикселей в изображении. Величина $p(z_i)$ является оценкой вероятности появления пикселя с интенсивностью z_i . Гистограммы различных фаций показаны на Рис. 1.

¹ Работа выполнена при частичной поддержке проекта 2.10 Программы ОНИТ РАН и Государственных контрактов № 02.740.11.0526, № 07.514.11.4048, предусматривающих обработку сигналов и изображений

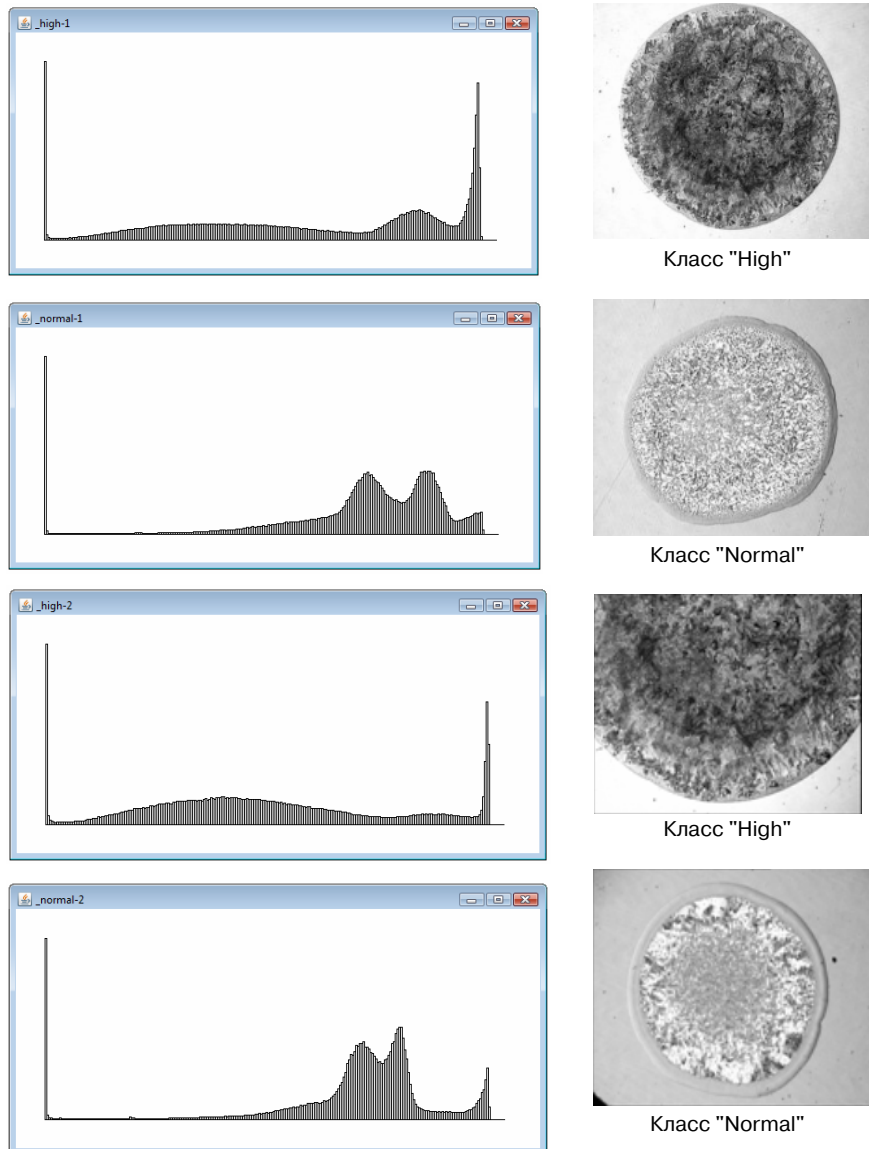


Рис. 1. Изображения фаций и гистограмм для двух классов заболевания

Для гистограмм выполняется условие нормировки: $\sum_{i=0}^{L-1} p(z_i) = 1$. Для дальнейшей работы конвертируем все цветные снимки в градации серого, используя NTSC фактор преобразования «цвет-яркость» [4]. Согласно стандарту ITU-R BT.601 яркость отдельного пикселя определяется по формуле: $z = 0.56G + 0.33R + 0.11B$, где R, G и B представляют собой компоненты вектора в пространстве RGB .

2. Удаление фона

Для постановки диагноза необходимо выделить на снимке очертания фации. При обработке изображения, получаемого с микроскопа, оператор либо размещает фацию в центре снимка, либо выделяет лишь фрагмент фации [5, 6]. Таким образом, получаем два типа изображений, поступающих на автоматическую обработку (Рис. 2).

На левом изображении фация полностью расположена на снимке, и пиксели, лежащие вдоль его границы, соответствуют предметному стеклу (подложке). Однако на рисунке справа фация занимает большую часть изображения, и пиксели фона находятся лишь в верхней его части. Как показывает визуальный анализ, внутренним точкам фации соответствуют меньшие значения уровней яркости.

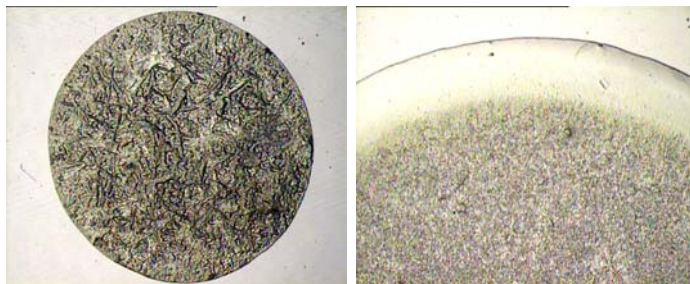


Рис.2 . Варианты расположения фации на снимке

Для удаления пикселей фона будем использовать алгоритм, основанный на применении гистограмм. На первом этапе работы алгоритма выделим области, расположенные по углам снимка (Рис. 3).

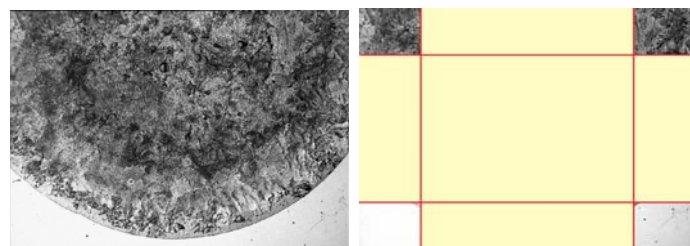


Рис. 3. Выделение областей на углах снимка

В одних случаях в эти области попадут части фации, в других — фрагменты фона. Построим для каждой выбранной области гистограммы интенсивностей яркостей (Рис. 4).

Для каждого из полученных распределений яркости найдем величину интенсивности в пике гистограммы $I_k = \arg \max p_k(z_i)$, где $k = 1, \dots, 4$ - номер выбранной области изображения [7].

Из всех найденных I_k найдем максимальное $I_{\max} = \max_k I_k$, которое и будет служить для определения интенсивностей пикселей, относящихся к фону. Для удаления областей фона

будем применять алгоритм, последовательно закрашивающий те пиксели изображения, которые удовлетворяют следующему условию:

$$I_{\max} - I_l < I(x, y) < I_{\max} + I_r,$$

где $I(x, y)$ – значение интенсивности пикселя в точке (x, y) , I_l и I_r – протяженность области фона слева и справа от пика [8].

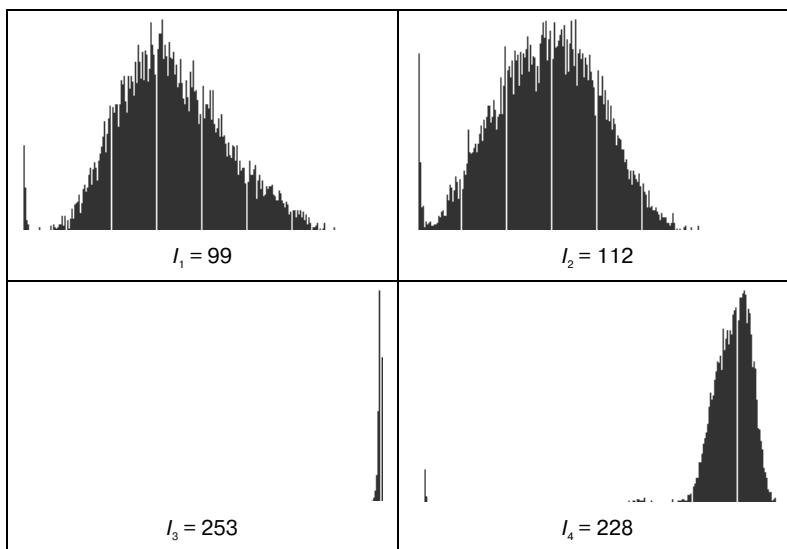


Рис. 4. Распределение яркостей по краям снимка фации

3. Нахождение дескрипторов для анализа изображения

Исходные изображения фаций (Приложение 1), участвующих в эксперименте, представлены в формате JPEG и цветовой модели *RGB* и классифицированы врачом-экспертом. После приведения снимков к полутоновому виду следует выделить признаки (дескрипторы), по которым можно проводить автоматическую классификацию. Важно отметить, что рост числа признаков увеличивает время, необходимое на обработку. Некоторые компоненты векторов признаков не несут новой информации об изображении и приводят к снижению значимости каждого отдельного признака. По этой причине возникает необходимость в поиске оптимального соотношения между качеством анализа и длиной вектора признаков. Рассмотрим ряд дескрипторов и оценим их значимость в задаче определения мочекаменной болезни. При этом ограничимся ее двумя классами: «High» (высокая степень заболевания) и «Normal» (отсутствие заболевания).

Математическое ожидание (МО) случайной величины z_i определяется формулой:

$$MO = \sum_{i=0}^{L-1} z_i p(z_i).$$

В Табл. 1 представлены результаты расчетов МО для изображений фаций, приведенных в Приложении 1.

Из таблицы видно, что изображения класса «High» характеризуются меньшими значениями параметра МО в отличие от снимков, относящихся к классу «Normal». Математическое ожидание дает достаточно грубое представление об интенсивности изображения. Для получения более точных показателей найдем стандартное отклонение от среднего уровня яркости.

Часто при описании изображения используются центральные моменты. Второй центральный момент $\mu_2(z) = \sigma^2(z)$ вычисляют по формуле:

$$\mu_2(z) = \sigma^2(z) = \sum_{i=0}^{L-1} (z_i - m)^2 p(z_i).$$

Дисперсия $\sigma^2(z)$ отражает разброс распределения яркостей изображения вокруг среднего

Табл. 1. Значения математического ожидания величины z

Номер образца	Класс "High"	Номер образца	Класс "Normal"
1	151,96	8	179,05
2	122,08	9	172,63
3	146,37	10	172,65
4	165,57	11	179,92
5	118,27	12	199,96
6	125,28	13	186,02
7	183,07	14	176,34
		15	170,83

Табл. 2. Значения дисперсии случайной величины z

Номер образца	Класс «High»	Номер образца	Класс «Normal»
1	5768,92	8	2729,19
2	4520,87	9	2413,11
3	5180,58	10	2524,17
4	341,06	11	3474,14
5	5534,65	12	3132,18
6	4720,17	13	2212,78
7	3689,59	14	2617,16
		15	2671,35
		15	7,088

значения. В Табл. 2 представлены значения второго центрального момента изображений различных классов.

На основе найденных значений дисперсии построим дескриптор относительной гладкости, нормировав его до интервала $[0, 1]$:

$$R = 1 - \frac{1}{1 + \frac{\sigma^2(z)}{(L-1)^2}}.$$

Из результатов, представленные в Табл. 2, можно заключить, что текстура изображений класса "Normal" характеризуется меньшей изменчивостью, т.е. является более «гладкой».

Однородность изображения задается следующим выражением:

$$U = \sum_{i=0}^{L-1} p^2(z_i).$$

Значение величины U для изображения уменьшается по мере роста яркостных различий. Из Табл. 3 видно, что снимки, относящиеся к классу высокой степени заболевания, отличаются большей равномерностью текстуры.

Этот факт объясняется тем, что в процессе камнеобразования белки «тянут» за собой соли в краевую (белковую) зону, поэтому изображенные фации при переходе от центральной к граничной зоне становится более однородным.

Энтропия служит для характеристики варибельности яркости и задается следующим выражением:

$$e = - \sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i).$$

Энтропия характеризует изменчивость яркости изображения. Для изображения, имеющего $p(z_i) = const$ для всех значений z_i , энтропия будет принимать наибольшее значение. В Табл. 4 представлены результаты вычислений для 15 исходных снимков фаций.

Максимальное значение энтропии показало шестое изображение, относящееся к классу «High» и характеризующееся практически полным отсутствием краевой белковой зоны. Признаки, полученные на основе статистики первого порядка, дают информацию, связанную с распределением уровней яркости, но из них нельзя получить информацию о взаимном расположении этих уровней в пределах изображения.

Матрица совместной встречаемости представляет собой оценку плотности распределения вероятности второго порядка $p_2(z_i, z_j)$ [9], полученную по одному изображению в предположении, что плотность вероятности зависит лишь от взаимного расположения пикселей с яркостями z_i и z_j . Обозначим матрицу через $C_r = (c_{ij})$, где r – отношение, в котором находятся пиксели i и j :

$$C_r = \frac{1}{s} \sum_{z_i, z_j: r} p_2(z_i, z_j),$$

а величина s соответствует числу сочетаний элементов, состоящих в отношении r .

Все диагональные элементы матрицы совместной встречаемости c_{ii} равны площадям соответствующих областей изображения, значения яркости которых равны z_i . Элементы матрицы c_{ij} ($i \neq j$), находящиеся вне главной диагонали,

Табл. 3. Характеристики равномерности текстуры изображений

Номер образца	Класс "High"	Номер образца	Класс "Normal"
1	0,014	8	0,061
2	0,011	9	0,13
3	0,014	10	0,053
4	0,057	11	0,068
5	0,01	12	0,204
6	,0060	13	0,464
7	0,026	14	0,191
		15	0,023

Табл. 4. Значения энтропии

Номер образца	Класс "High"	Номер образца	Класс "Normal"
1	7,239	8	6,607
2	7,293	9	5,886
3	7,458	10	6,582
4	6,535	11	6,368
5	6,916	12	5,222
6	7,525	13	3,268
7	7,207	14	5,114
		15	7,088

равны длинам границ, разделяющих соответствующие области изображения, которые образованы пикселями с яркостями z_i и z_j .

Отношение r определяется с помощью расстояния d и угла θ . Для нашей задачи будем рассматривать пиксели со значением $d=1$ и значением угла θ , равного $0^\circ, 45^\circ, 90^\circ$ и 135° .

Такие значения параметров соответствуют отношению ближайших соседей.

Будем вычислять значения элементов матрицы C_r для каждого снимка фации по следующему алгоритму:

```

Begin
  Установка нулевых значений в массиве  $C_r[i, j]$ 
  for (<все значения  $i$ >)
    for (<все значения  $j$ >)
      if (< $z_i$  и  $z_j$  состоят отношении  $r$ >)
         $C_r[z_i, z_j]++$ ;
  for (<все  $i$  и  $j$ >)
     $C_r[i, j] /= s$ ;
End
    
```

Анализ матрицы C_r позволяет определить, к какой категории относится текстура области. Для этой цели предлагается использовать следующий набор дескрипторов:

1. Максимум вероятности

$$p_{\max} = \max_{i,j}(c_{ij}).$$

Данный дескриптор указывает значение наиболее сильного отклика на отношение r .

2. Однородность для матрицы C_r :

$$U_c = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} c_{ij}^2.$$

3. Средняя энтропия:

$$e_c = - \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} c_{ij} \log_2 c_{ij}.$$

В Табл. 5 представлены значения вышеперечисленных признаков для каждого из исходных образцов фаций.

4. Классификация изображений фаций на основе статистического анализа

Рассмотрим бинарную задачу распознавания мочекаменной болезни с учетом выделенных семи дескрипторов. Для измерения расстояний до каждого из классов используем метрики Евклида и Махаланобиса [10]. В Приложении 2 представлены расстояния Евклида между имеющимися образцами фаций. Построим минимальное покрывающее дерево. Согласно алгоритму Краскала [11], по таблице расстояний строится список ребер графа и упорядочивается в порядке возрастания длин. Из введенного списка последовательно извлекаются ребра. Если выбранное ребро не приводит к возникновению цикла, то оно включается в дерево. Если в результате выполнения алгоритма образуется более одного дерева, то существует ребро, при добавлении которого не возникает цикла – оно соединяет два дерева.

На Рис. 5 изображено полученное экспертами остовное дерево. Объекты 1-7 изначально относятся к первому классу («High»), а 8-15 – ко второму («Normal»).

В Табл. 6 приведены значения кратчайших расстояний между образцами. В то же время видно, что образец под номером 11 классифицируется построенным деревом неверно.

Табл. 5. Признаки второго рода для исходных изображений

Номер образца	p_{\max}	U_c	e_c
1	0,041	0,0020	12,992
2	0,04	0,0020	13,629
3	0,042	0,0020	13,665
4	0,038	0,0020	13,013
5	0,059	0,0040	12,547
6	0,129	0,017	13,196
7	0,037	0,0020	13,301
8	0,039	0,0020	12,730
9	0,039	0,0020	11,736
10	0,038	0,0020	13,054
11	0,043	0,0050	12,173
12	0,049	0,0070	10,917
13	0,039	0,0020	11,152
14	0,039	0,0020	11,947
15	0,038	0,0020	13,353

Табл. 6. Расстояния между ближайшими образцами

Номер образца	Класс	Номер ближайшего образца	Расстояние Евклида
1	High	3	588,367
2	High	3	660,158
3	High	1	588,367
4	High	11	62,746
5	High	1	2821680,010
6	High	2	33920,586
7	High	11	215,486
8	Normal	15	58,434
9	Normal	10	111,066
10	Normal	14	93,074
11	Normal	4	62,746
12	Normal	4	28,986
13	Normal	9	200,780
14	Normal	15	54,485
15	Normal	14	54,485

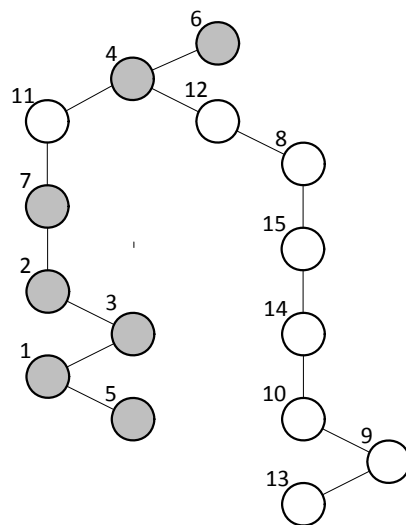


Рис. 5. Остовное дерево минимального веса

Табл. 7. Матрица ковариаций для образцов класса "High"

36,958	663,99	-0,183	2,081	0,081	0,014	1,006
663,99	553531,031	54,022	-476,437	-12,197	-1,834	82,371
-0,183	54,022	0,0080	-0,076	-0,0010	0,0	-0,0020
2,081	-476,437	-0,076	0,709	0,015	0,0020	0,04
0,081	-12,197	-0,0010	0,015	0,0010	0,0	-0,0020
0,014	-1,834	0,0	0,0020	0,0	0,0	0,0
1,006	82,371	-0,0020	0,04	-0,0020	0,0	0,153

Табл. 8. Матрица ковариаций для образцов класса "Normal"

264,755	12373,525	-2,943	18,647	-0,021	-0,01	13,043
12373,525	621935,504	-140,822	879,515	-0,825	-0,414	614,641
-2,943	-140,822	0,034	-0,216	0,0	0,0	-0,142
18,647	879,515	-0,216	1,369	-0,0010	-0,0010	0,891
-0,021	-0,825	0,0	-0,0010	0,0	0,0	-0,0020
-0,01	-0,414	0,0	-0,0010	0,0	0,0	-0,0010
13,043	614,641	-0,142	0,891	-0,0020	-0,0010	0,761

В отличие от метрики Евклида, расстояние Махаланобиса учитывает корреляцию между компонентами векторов (признаками) и определяется по следующей формуле:

$$d_M = \sqrt{(x - y)^T C^{-1} (x - y)},$$

где C – ковариационная матрица [10], составленная из парных ковариаций элементов векторов $x = (x_1, \dots, x_7)^T$ и $y = (y_1, \dots, y_7)^T$. Данная матрица представляет собой математическое ожидание произведения центрированных случайных величин:

$$C = \text{cov}(x, y) = M[(x - Mx)(y - My)],$$

где Mx – математическое ожидание случайного вектора x .

Результаты расчетов отражены в Табл. 7 и Табл. 8.

Для измерения расстояний были использованы расчетные значения математических ожиданий дескрипторов:

- для образцов первого класса

$Mx = (48,072; 1335,437; 0,078; 5,928; 0,055; 0,0040; 13,192);$

- для образцов второго класса

$Mx = (33,345; 2166,232; 0,418; 3,391; 0,04; 0,0030; 12,133).$

Экспериментально установлено, что применение метрики Махаланобиса для измерения

расстояний между образцами и сформированными классами приводит к правильной бинарной классификации фаций, что определяет целесообразность ее использования для обнаружения мочекаменной болезни. Развитием предложенного подхода может стать установление степени заболевания.

Заключение

В работе рассмотрен метод бинарной классификации изображений на основе предварительного выделения статистических признаков (дескрипторов). Предложен алгоритм удаления фона, что позволило повысить качество обработки изображений и точность определения признаков. В рамках задачи классификации мочекаменной болезни исследована эффективность применения метрик Евклида и Махаланобиса, показавшая преимущество последней. Для проведения экспериментов создано программное обеспечение, реализующее алгоритмы предварительной обработки снимков, формирования дескрипторов и классификации изображений. В дальнейшем, с целью повышения эффективности метода, предполагается переход на параллельную обработку потоков фаций с использованием высокопроизводительных кластерных установок, обеспечивающих масштабирование времени вычисления с ростом числа вычислительных узлов.

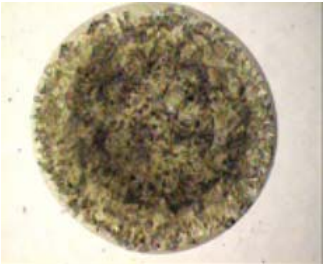
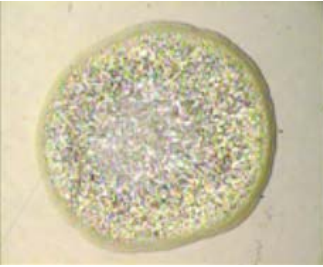
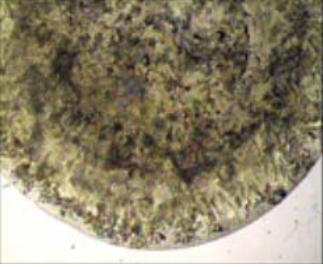

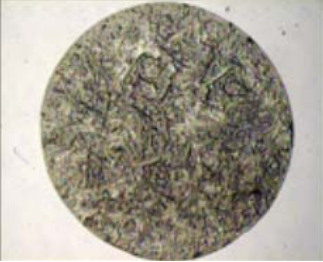
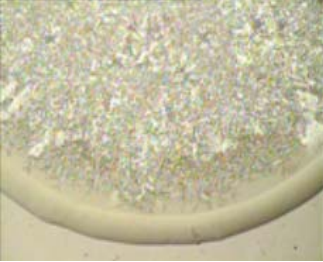
Литература

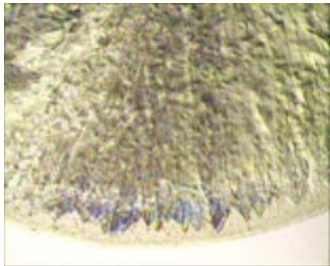
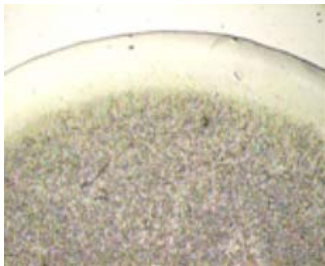

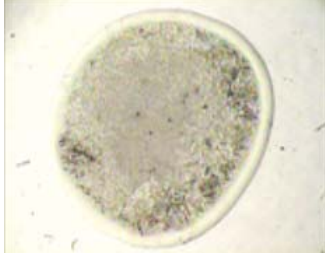
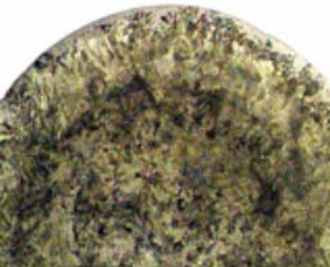

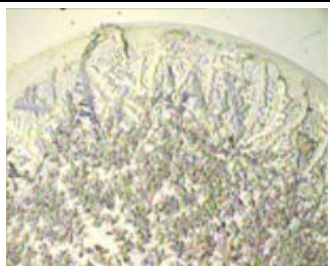

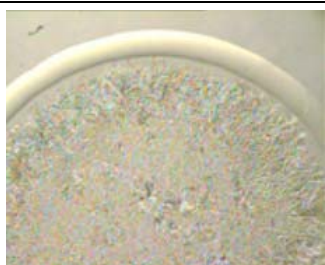
1. Дасаева Л.А., Шабалин В.Н., Шатохина С.Н., Шилов Е.М. Новое в диагностике мочекаменной болезни – <http://www.medtsu.tula.ru/VNMT/Archive/2004/n3/p47.htm>
2. Theodoridis S., Koutroumbas K. Pattern Recognition – Elsevier Academic, 2003. <http://en.wikipedia.org/wiki/NTSC>
3. Павлидис Т. Алгоритмы машинной графики и обработки изображений. – М.: Радио и связь, 1991.
4. Сойфер В.А. Методы компьютерной обработки изображений – М.: Физматлит, 2003.
5. Лисовая Н.А. Лабораторные подходы к выявлению мочекаменной болезни – С-Пб: Terra Medica.
6. Гонсалес Р., Вудс Р. Цифровая обработка изображений – М.: Техносфера, 2006.
7. Виноградов А.Н. Калугин Ф.В. Недев М.Д. Погодин С.В. Талалаев А.А. Тищенко И.П. Фраленко В.П. Хачумов В.М. Выделение и распознавание локальных объектов на аэрокосмических снимках. – М.: Авиакосмическое приборостроение, № 9, 2007, 39-45 с.
8. MacKay David. Information Theory, Inference and Learning Algorithms. – Cambridge University Press, 2003.
9. Новиков Ф.А. Дискретная математика для программистов – С-Пб: Питер, 2008.
10. Толмачев И.Л., Хачумов М.В. Бинарная классификация на основе варьирования размерности пространства признаков и выбора эффективной метрики. – Искусственный интеллект и принятие решений, № 2, 2010, с.3-10.
11. Бочаров П. П., Печинкин А. В. Теория вероятностей. Математическая статистика – М.: Физматлит, 2005.

Мажуга Вера Владимировна. Магистрант Российского университета дружбы народов (РУДН). Автор двух печатных работ. Область научных интересов: интеллектуальный анализ данных, информационные технологии. E-mail: vvmazhuga@inbox.ru

Хачумов Михаил Вячеславович. Аспирант Российского университета дружбы народов (РУДН). Автор 13-ти работ. Область научных интересов: классификация и кластеризация информации. E-mail: khmike@inbox.ru

Приложение 1. Исходные изображения фаций

Класс «High»		Класс «Normal»	
1		8	
2		9	
3		10	

4		11	
5		12	
6		13	
7		14	
	15		

Приложение 2. Матрица расстояний Евклида между исходными образцами

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.0	1248.409	588.367	2355.898	2821680.01	33936.22	2079.555	3039.843	3355.874	3244.818	2294.952	2637.176	3556.306	3151.853	3097.629
2	1248.409	0.0	660.158	1108.661	2821680.182	33920.586	833.506	1792.578	2108.366	1997.342	1048.329	1390.872	2308.979	1904.481	1850.163
3	588.367	660.158	0.0	1767.623	2821680.022	33923.131	1491.433	2451.601	2767.595	2656.542	1706.772	2049.101	2968.068	2563.594	2509.35
4	2355.898	1108.661	1767.623	0.0	2821680.798	33945.199	277.09	683.997	999.977	888.922	62.746	282.986	1200.46	795.973	741.731
5	2821680.01	2821680.182	2821680.022	2821680.798	0.0	2855600.116	2821680.604	2821681.395	2821681.727	2821681.606	2821680.753	2821681.024	2821681.956	2821681.509	2821681.453
6	33936.22	33920.586	33923.131	33945.199	2855600.116	0.0	33935.701	33978.423	33998.399	33991.044	33942.922	33957.233	34012.602	33985.168	33981.85
7	2079.555	833.506	1491.433	277.09	2821680.604	33935.701	0.0	960.409	1276.532	1165.477	215.486	557.678	1476.825	1072.459	1018.323
8	3039.843	1792.578	2451.601	683.997	2821681.395	33978.423	960.409	0.0	316.154	205.13	744.942	403.531	516.47	112.072	58.434
9	3355.874	2108.366	2767.595	999.977	2821681.727	33998.399	1276.532	316.154	0.0	111.066	1061.054	719.592	200.78	204.086	258.25
10	3244.818	1997.342	2656.542	888.922	2821681.606	33991.044	1165.477	205.13	111.066	0.0	949.999	608.631	311.684	93.074	147.192
11	2294.952	1048.329	1706.772	62.746	2821680.753	33942.922	215.486	744.942	1061.054	949.999	0.0	342.546	1261.376	856.984	802.842
12	2637.176	1390.872	2049.101	282.986	2821681.024	33957.233	557.678	403.531	719.592	608.631	342.546	0.0	919.51	515.563	461.76
13	3556.306	2308.979	2968.068	1200.46	2821681.956	34012.602	1476.825	516.47	200.78	311.684	1261.376	919.51	0.0	404.501	458.828
14	3151.853	1904.481	2563.594	795.973	2821681.509	33985.168	1072.459	112.072	204.086	93.074	856.984	515.563	404.501	0.0	54.485
15	3097.629	1850.163	2509.35	741.731	2821681.453	33981.85	1018.323	58.434	258.25	147.192	802.842	461.76	458.828	54.485	0.0