

УДК 681.007:61.1

МЕТОД РАСПОЗНАВАНИЯ ОБЪЕКТОВ С ДИНАМИЧЕСКИМИ ХАРАКТЕРИСТИКАМИ

М. К. Яковенко,

аспирант

Санкт-Петербургский государственный университет
аэрокосмического приборостроения (ГУАП)

Распознавание объектов с динамическими характеристиками — сложная задача, обусловленная высокой вариабельностью их ключевых параметров. Основная проблема — это выявление эффективного множества признаков, способных описать объект в достаточной мере для корректного распознавания с приемлемым уровнем. В данной статье в качестве примера объектов с динамическими характеристиками рассмотрены лейкоциты человека (пять основных классов).

Recognition of objects with dynamic characteristics is hard task because of very high variability of their key parameters. Main problem is extraction of effective set of features that will describe target objects enough for classification with acceptable recognition rate. In this paper is considered good example of objects with dynamic characteristics — human white blood cell (five main classes).

Введение

Распознавание объектов с динамическими характеристиками является далеко не тривиальной задачей не только для искусственного интеллекта, но и для эксперта. В качестве яркого примера подобных объектов можно отметить биологические материалы живых организмов.

Тестовыми объектами для разработки описываемого в настоящей работе метода послужили лейкоциты человека — одни из наиболее сложных для классификации биологических объектов ввиду их значительной вариабельности. Помимо разных этапов жизненного цикла самих клеток (они могут быть молодыми, зрелыми, отработанными), на вид лейкоцитов в значительной мере оказывает влияние физическое состояние и здоровье человека. Также сказываются экологические условия, вредные воздействия (например, радиоактивное) и даже время суток, в которое был взят анализ крови.

Актуальность проблемы создания автоматического классификатора лейкоцитов основывается на необходимости проведения в повседневной жизни огромного количества рутинных исследований с привлечением высококвалифицированного персонала.

Описание предметной области

В качестве анализируемых объектов в данной работе использованы основные лейкоциты периферической крови иммунной системы человека.

Это пять типов клеток: нейтрофил, эозинофил, базофил, лимфоцит и моноцит (рис. 1 — см. на 3-й стороне обложки).

Клетки распознаются врачом-гематологом по совокупности характеристик, описанных в работе [1]. Следует отметить высокую вариабельность размеров, форм и окрасок ядра и цитоплазмы каждого класса клеток. Медицинский алгоритм распознавания клеток изобилует такими терминами как зернистость, глыбчатость, сетчатость, гранулированность, равномерность, размытость. Все это описания *текстуры* ядер и цитоплазм клеток. Текстура является одним из важнейших признаков, по которым опытный гематолог с высокой вероятностью классифицирует клетки крови.

Изображения лейкоцитов для проведения исследований были получены из публичной базы данных [2], предназначенной для обучения и повышения квалификации врачей. Эта база составлена из мазков, взятых в различное время в нескольких госпиталях Швеции. Мазки принадлежат людям разного возраста и обоих полов. Лейкоциты классифицированы высококвалифицированными профессионалами-гематологами. Все клетки окрашены по методу May-Grunwald-Giemsa. Для экспериментов предоставлено по 50 клеток каждого класса.

Выбор признаков описания объектов

Из описания предметной области лейкоцитов, с учетом опыта профессиональных врачей-гематологов, был сделан вывод о высокой дискрими-

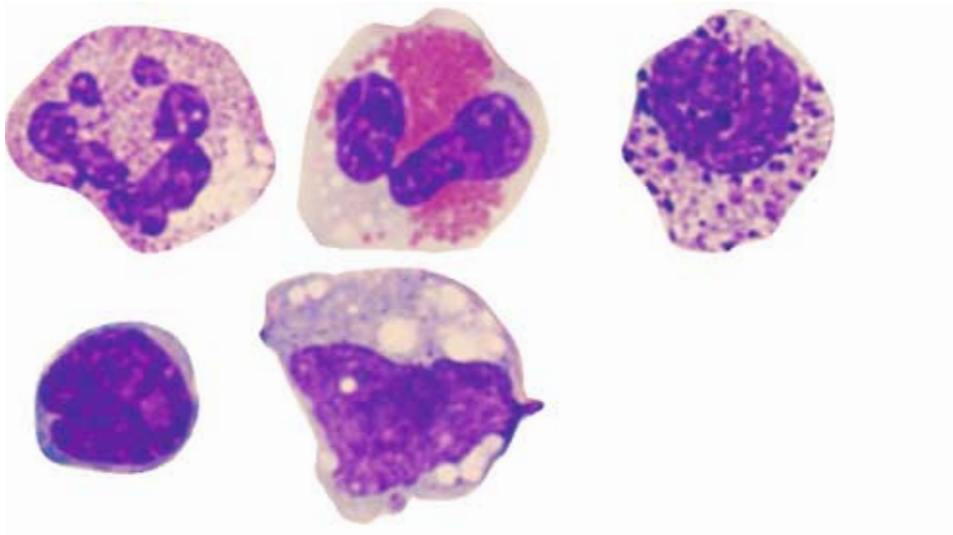


Рис.1. Объекты исследования: нейтрофил, эозинофил, базофил, лимфоцит и моноцит

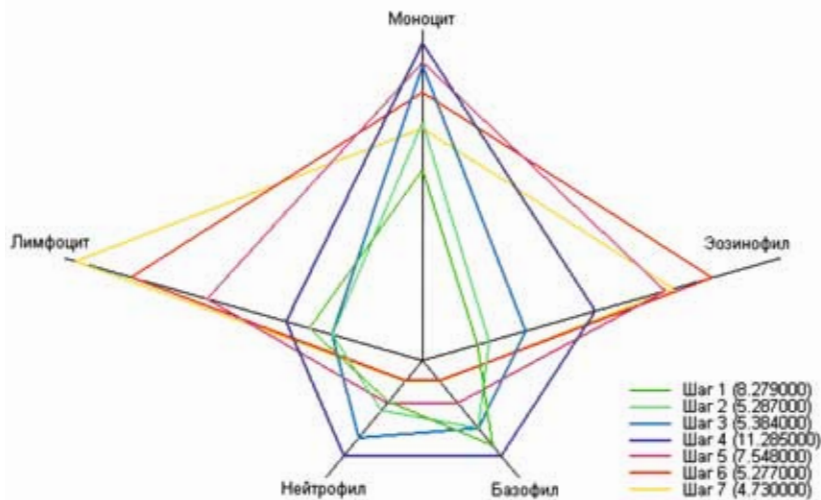


Рис.4. График для определения оптимального шага построения МСВУЯ

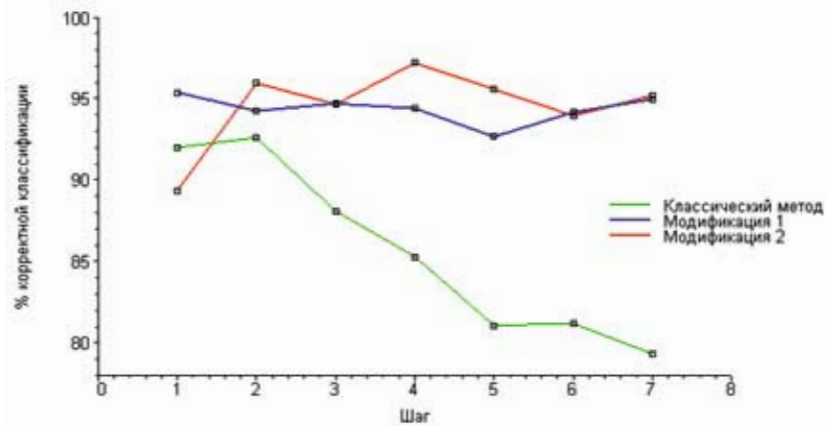


Рис.5. Сводный график результативности различных методов формирования МСВУЯ (показан средний по классам процент корректной классификации)

национной способности текстуры клеток. Следует отметить, что применение текстурных описателей при составлении пространства признаков объектов целесообразно не только в случае лейкоцитов. Исследователи целого ряда природных и, в частности, биологических объектов отмечают высокий потенциал текстуры.

В литературе рассмотрено большое количество различных методов описания текстуры. На данный момент не существует единого мнения о том, какие текстурные дескрипторы наиболее эффективны. Исследователи отдают предпочтение тем или иным методам. При анализе текстурных описателей применяются либо изображения из специальных тестовых баз данных текстур, либо искусственно созданные текстуры с теми или иными характеристиками, либо текстуры реальных объектов исследования.

В [3] производится сравнительный анализ четырех подходов к описанию текстуры: 1) анализ спектра Фурье; 2) разности уровней яркости (РУР); 3) матрицы совместной встречаемости уровней яркости (МСВУЯ); 4) статистика длин примитивов (СДП). Авторы приходят к выводу, что статистические текстурные описатели лучше фиксируют сущность текстуры, нежели подход с использованием спектра Фурье.

В [4] проводятся исследования тех же методов описания текстуры, что и у предыдущих авторов. Тестирование осуществляется на сгенерированных текстурах на основе случайных полей второго порядка и полей Маркова. Наиболее мощным признан метод МСВУЯ.

В [5] сравниваются фрактальный подход, МСВУЯ и распределение энергии текстуры (РЭТ). По итогам исследований последние два метода дали лучшие результаты.

В [6] анализируется производительность четырех подходов к получению текстурных дескрипторов: 1) марковские случайные поля (МСП); 2) фильтр Габор; 3) фрактальный подход; 4) МСВУЯ. Для тестирования использованы фрактальные изображения, изображения кожи, рисованные картины, изображения со свойствами гауссовых МСП. Лучшими описателями признаны МСВУЯ и фракталы. Авторы отмечают, что МСВУЯ эффективнее применять в маленьких окнах, а не по всему изображению. Также показано, что использование нескольких направлений при формировании МСВУЯ увеличивает вероятность корректной классификации.

В [7] сравниваются несколько методов для классификации спутниковых изображений с использованием нейронных классификаторов. Проанализированы следующие подходы: МСВУЯ, РУР, фильтр Габор, анализ спектра Фурье. Автор делает вывод, что универсальный набор дескрипторов, показывающий наилучшие результаты на изображениях различных классов, построить невозможно. В целом предпочтение автора было отдано спектру Фурье, второе место поделили МСВУЯ и фильтр Габор.

Из приведенного анализа литературных источников можно констатировать превосходство МСВУЯ

над прочими методами построения текстурных описателей. Однако следует отметить, что у исследователей встречаются и прямо противоположные результаты. В [8] приводится сравнительный анализ следующих методов: РЭТ, фильтр Габор, вэйвлеты, спектр Фурье, МСВУЯ. Различные методы были лучшими на разных множествах текстурных изображений. При этом МСВУЯ оказались худшими во всех случаях, а предпочтение отдано вэйвлетам. Также в пользу вэйвлетов высказываются исследователи в [9]. В данном анализе помимо вэйвлетов использовались спектр Фурье и фильтр Габор. Последний метод наравне с вэйвлетами показал хорошие результаты. Интересно отметить, что исследования, проведенные авторами в [6], были повторены, а результаты — опровергнуты другим исследователем [10]. В его работе наилучшую производительность показали МСП, а худшую — МСВУЯ.

Таким образом, по всей видимости, эффективность текстурных дескрипторов во многом зависит от предметной области, объектов исследования. Также следует отметить, что не все результаты исследований можно принимать с большой достоверностью, так как среди них встречаются неглубокие проработки того или иного подхода. Например, исследователи используют в своих тестах ограниченный набор текстурных дескрипторов, которые можно получить по определенному методу. Также может оказаться недостаточной или вовсе отсутствовать оптимизация параметров метода.

Проведенные исследования и эксперименты с МСВУЯ показали достаточно хорошие результаты при применении на изображениях клеток. Далее приводится описание метода МСВУЯ.

Положим размер анализируемого изображения на $N_x \times N_y$, число уровней яркости — N_g . Пусть $L_x = \{1, 2, \dots, N_x\}$ и $L_y = \{1, 2, \dots, N_y\}$ — соответственно горизонтальная и вертикальная пространственная область, $G = \{1, 2, \dots, N_g\}$ — множество уровней яркости изображения. Тогда изображение может быть определено следующим образом: $I : L_x \times L_y \rightarrow G$.

МСВУЯ представляют собой оценку плотности распределения вероятностей второго порядка. Элемент матрицы — условная вероятность $P(i|j)$ возникновения уровня яркости i на расстоянии d от уровня яркости j . Поскольку число таких матриц может быть очень большим, необходимы некоторые упрощения. Имеет смысл высчитывать четыре матрицы для заданного шага d для $0^\circ, 45^\circ, 90^\circ$ и 135° расположения пар уровней яркости. Формально элементы матриц для указанных направлений могут быть записаны как

$$P(i|j)_0 = \#\{(k, l), (m, n) \in (L_x \times L_y) | (k - m = 0, |l - n| = d), \\ l(k, l) = i, l(m, n) = j\};$$

$$P(i|j)_{45} = \#\{(k, l), (m, n) \in (L_x \times L_y) | (k - m = d, |l - n| = d), \\ l(k, l) = i, l(m, n) = j\};$$

$$P(i|j)_{45} = \#\{(k, l), (m, n) \in (L_x \times L_y) \mid (k-m=d, l-n=-d) \vee (k-m=-d, l-n=d), l(k, l)=i, l(m, n)=j\}$$

$$P(i|j)_{90} = \#\{(k, l), (m, n) \in (L_x \times L_y) \mid (|k-m|=d, l-n=0), l(k, l)=i, l(m, n)=j\}$$

$$P(i|j)_{135} = \#\{(k, l), (m, n) \in (L_x \times L_y) \mid (k-m=d, l-n=d) \vee (k-m=-d, l-n=-d), l(k, l)=i, l(m, n)=j\}$$

где знак # обозначает количество элементов в множестве.

Полученные матрицы являются квадратными с размером, равным N_g . Также они являются симметричными, т. е. $P(i|j) = P(j|i)$.

Можно провести следующую нормализацию матриц для оперирования с вероятностями:

$$N(i|j) = \frac{P(i|j)}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P(i|j)}$$

МСВУЯ самостоятельно обычно не используются в качестве текстурных описателей. На их основе строятся разнообразные характеристики [3, 4, 11, 12]. В данной работе используется 24 признака. Для примера рассмотрим три из них:

$$T_1 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (P(i|j))^2; \tag{1}$$

$$T_2 = \sum_{n=1}^{N_g-1} n^2 \left(\sum_{\substack{i=1 \\ |i-j|=n}}^{N_g} P(i|j) \right); \tag{2}$$

$$T_3 = \left(\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (ij)P(i|j) - \mu_x \mu_y \right) / (\sigma_x \sigma_y), \tag{3}$$

где μ_x, μ_y — и математические ожидания частных распределений, связанных с $P(i|j)$ по строкам и столбцам матриц соответственно; а σ_x, σ_y — соответствующие стандартные отклонения.

Выражение (1) описывает угловой второй момент, являющийся мерой однородности текстуры. Формула (2) определяет контраст изображения. Контраст является дифференциальным моментом матрицы совместной встречаемости и дает количественную оценку локальных изменений яркости, присутствующих на изображении. Параметр, заданный по выражению (3), — это корреляция, являющаяся мерой присутствия полутоновой линейной зависимости на изображении.

Каждый из построенных текстурных дескрипторов на основе МСВУЯ должен быть вычислен четыре раза — для каждого направления ($0^\circ, 45^\circ, 90^\circ$ и 135°). При этом теоретически каждый дескриптор из группы наиболее точно отражает лишь свое направление. Исследования [11] показали, что мож-

но использовать среднее значение в каждой группе дескрипторов в качестве текстурного описателя, инвариантного по вращению. Дополнительно к среднему рекомендуется использовать разброс в группе описателя (минимальное и максимальное значения).

Единственной сложностью остается выбор параметра шага d . Несмотря на то, что проведенные исследования и эксперименты с МСВУЯ показали достаточно хорошие результаты при применении на изображениях клеток, тем не менее, за счет корректного выбора шага d не удалось добиться высокой степени дифференцированности кластеров. Поэтому была проанализирована возможность модернизации МСВУЯ для повышения информативности получаемых из них текстурных дескрипторов.

МСВУЯ могут быть представлены в нотации дискретных цепей Маркова. Пусть система меняет свое состояние в моменты времени $t = 1, 2, \dots, n$. При этом она может находиться в одном из r состояний s_1, s_2, \dots, s_r , где $r > 2$. Пусть $A_k^{(j)}$ — событие того, что система на k -м шаге входит в состояние s_j . Тогда, если для всех $k = 2, \dots, n$ выполняется

$$P[A_k^{(jk)} | A_{k-1}^{(jk-1)}, \dots, A_1^{(j1)}] = P[A_k^{(jk)} | A_{k-1}^{(jk-1)}],$$

то данная система является цепью Маркова. Состояние такой системы в определенный момент времени зависит только от предыдущего состояния и не зависит от прошлой истории состояний системы. Таким образом, это система без памяти о прошлом. Обозначим $P(i|j)$ вероятность того, что цепь Маркова в момент времени t находится в состоянии s_i при условии, что в момент времени $t-1$ она находилась в состоянии s_j . Причем эта вероятность не зависит от конкретного значения t . Заданные таким образом вероятности являются вероятностями перехода и представляются в виде матрицы:

$$P = \begin{bmatrix} P(1|1) & \dots & P(1|r) \\ \vdots & \ddots & \vdots \\ P(r|1) & \dots & P(r|r) \end{bmatrix}$$

Данная матрица эквивалентна введенной выше матрице совместной встречаемости уровней яркости.

В случае шага, равного 1, метод МСВУЯ функционирует очевидным образом, как описано выше. При этом мы имеем симметричную матрицу. При увеличении шага в классическом варианте рассматриваются пары пикселей на расстоянии $d > 1$. В этом случае изменяются только параметры построения матриц, но информации о текстуре в них не добавляется. Видимо, с этим связано употребление шага, равного 1, подавляющим большинством исследователей при построении МСВУЯ.

В данной работе предлагается повысить информативность МСВУЯ при использовании шага $d > 1$, несколько модифицировав при этом классическое представление о взаимодействии элементов изображения.

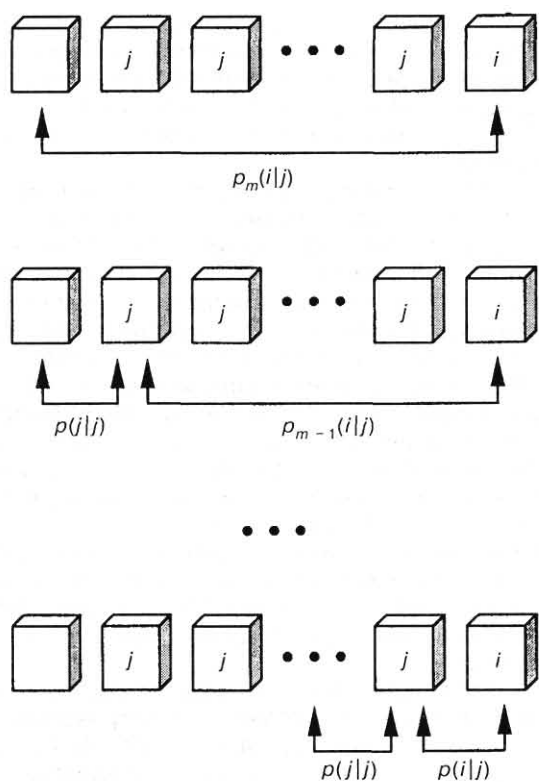


Рис. 2. Порядок вычисления вероятности перехода для формирования модифицированных МСВУЯ

В теории Маркова m -шаговая вероятность перехода определена [13] как

$$P_m = (i | j) = \sum_{k=1}^r P_{m-1}(i | k) P(k | j).$$

Это вероятность того, что цепь Маркова в момент времени t находится в состоянии s_i при условии, что в момент времени $t - m$ она находилась в состоянии s_j .

Для увеличения информативности МСВУЯ предлагается при вычислении вероятностей переходов в каждый момент времени t придерживаться следующих условий:

$$\begin{cases} p_m(i | j) \neq 0 \\ p_{m-1}(i | j)p(j | j) \neq 0 \\ \dots \\ p(i | j) \neq 0 \end{cases}$$

Таким образом, одна условная вероятность будет характеризовать переход из одного пикселя в группу пикселей изображения, т. е. под состоянием предлагается понимать не только один пиксель, но и последовательность пикселей одной яркости (рис. 2).

За счет применения такого подхода увеличивается количество информации, заложенной в переход между состояниями. При этом матрица пе-

рестает быть симметричной, что влечет за собой увеличение информативности сформированных на ее основе текстурных дескрипторов. Несимметричность матрицы также позволяет ввести дополнительные описатели текстуры, которые чаще всего не имели смысла при классическом способе формирования МСВУЯ, например, треугольную симметрию.

В рамках дальнейшего изучения возможности увеличения информативности текстурных признаков, построенных на базе МСВУЯ, рассмотрим марковские случайные поля. Как отмечалось выше, исследователи выделяют этот метод как имеющий высокий потенциал для описания текстур.

Теория случайных полей Маркова представляет математическую основу для решения проблемы получения информации о процессе в целом по имеющейся локальной информации. МСП являются двурным обобщением дискретных цепей Маркова. Обозначим $F = \{F_1, \dots, F_m\}$ множество случайных чисел, определенных на множестве позиций $S = \{S_1, \dots, S_m\}$. Набор значений $f = \{f_1, \dots, f_m\}$ называется конфигурацией F . Значения конфигурации принадлежат множеству возможных меток $L = \{L_1, \dots, L_k\}$. В случае изображений множество меток содержит возможные значения пикселей. Позиции взаимодействуют друг с другом в некотором окружении. Окружением позиции является $N = \{N_i | \forall i \in S\}$, где N_i — множество позиций, окружающих i . При этом позиция не входит в свое окружение: $i \in N_i$, и окружение имеет свойство обобщности: $i \in N_i \Leftrightarrow i' \in N_i$. В случае двумерного изображения множество элементов окружения радиуса r определяется выражением

$$N_i = \{i' \in S | dist(S_i, S_{i'})^2 \leq r, i \neq i'\},$$

где $dist(A, B)$ — Евклидово расстояние между точками.

Количество анализируемых «соседей» определяется порядком модели МСП. На рис. 3 представлена структура используемых множеств «соседей» для моделей до седьмого порядка включительно [14].

		7	6	7		
	5	4	3	4	5	
7	4	2	1	2	4	7
6	3	1	<i>i</i>	1	3	6
7	4	2	1	2	4	7
	5	4	3	4	5	
		7	6	7		

Рис. 3. Структура окружения для разных порядков моделей МСП

Совместная вероятность возникновения определенной конфигурации обозначается как $P(F_1 = f_1, \dots, F_{m1} = f_m) = P(f)$. F образует МСП на S относительно системы окружения N тогда и только тогда, когда одновременно выполняются два условия:

- 1) $P(f) > 0, \forall f \in F$ — позитивность;
- 2) $P(f_i | f_{S-(i)}) = P(f_i | f_{N_i})$ — свойство Маркова, где $f_{N_i} = \{f_{i'} | i' \in N_i\}$.

Условие позитивности необходимо в силу некоторых технических причин и обычно может быть выполнено на практике. Например, когда выполняется это условие, совместная вероятность $P(f)$ любого случайного поля уникально определяется через локальные условные вероятности [15]. Свойство Маркова означает, что конфигурация позиции зависит только от ближайшего окружения. Всегда можно выбрать размер окружения N_i такой, что это свойство будет достигнуто. Окружение более высокого порядка включает в себя все позиции окружений более низких порядков.

В [15] доказывается тот факт, что совместная вероятность $P(F_1 = f_1, \dots, F_{m1} = f_m) = P(f)$ имеет функцию распределения Гиббса:

$$P(f) = \frac{e^{-\frac{1}{T}U(f)}}{Z},$$

где $Z = \sum_{f \in F} e^{-\frac{1}{T}U(f)}$ — нормализационная константа,

так называемая термодинамическая сумма; T — глобальный контролирующий параметр, так называемая температура, которая обычно выбирается равной 1; $U(f)$ — функция энергии, характеризующая текстуру.

Существует несколько подходов к построению функции $U(f)$ по имеющемуся классу текстур [14]. Взаимосвязь МСП и распределения Гиббса описывается теоремой Hammersley—Clifford [15]. Ее практическая польза состоит в возможности задания совместной вероятности для изображения. Таким образом, МСП характеризуется свойством локальности, и в то же время имеет глобальное свойство (распределение Гиббса).

Если положить нормальный закон распределения появления уровней яркости в каждой позиции (μ, σ) , то мы будем иметь гауссовское МСП. Так как распределение Гаусса принадлежит семейству распределений Гиббса, то в этом случае распределение совместной вероятности примет вид

$$P(f) = \frac{\sqrt{|B|}}{\sqrt{(2\pi\sigma^2)^m}} e^{-\frac{(f-\mu)^T B(f-\mu)}{2\sigma^2}},$$

где m — $1 \times m$ вектор из m_i ; B — $m \times m$ матрица взаимодействия, элемент которой имеет вид $b_{i,i'} = \delta_{i,i'} - \beta_{i,i'}$, где $\beta_{i,i'}$ — константа, отражающая взаимодействие пары положений i и i' .

С использованием модели гауссовского МСП в [16] были получены признаки описания текстур

практически без потерь информации. С их помощью удавалось синтезировать копии текстур, близкие к оригиналам. Это является практической иллюстрацией эффективности использования модели МСП для выявления характерных свойств изображения.

Для второй модификации способа построения МСВУЯ использованы приведенные выше доказательства высокой информативности локального окружения модели МСП для описания текстур. При этом для формирования статистики совместной встречаемости применяется окружение, описываемое в модели. Шаг построения МСВУЯ ассоциируется в данном случае с порядком модели МСП.

При таком методе формирования МСВУЯ строится только одна матрица, учитывающая все направления текстуры одновременно. Для каждого положения i определяется набор условных вероятностей в виде $\{P(f_i | f_k), \forall k \in N_i\}$.

Теоретические выкладки, описывающие представленные выше методы модификации классического подхода формирования МСВУЯ, нашли адекватное отражение в практических результатах экспериментов.

Для уменьшения вычислительной нагрузки при работе с матрицами совместной встречаемости имеет смысл понизить их размерность. Для этого необходимо квантовать изображение на меньшее количество уровней яркости. Классическое яркое изображение имеет 256 градаций серого. На практике применения МСВУЯ оказывается достаточно восьми [11]. Для корректного понижения количества уровней яркости может быть использована процедура равновероятностной квантизации.

Оптимизация параметров признаков

Матрицы совместной встречаемости в классической нотации имеют два параметра: направление и шаг. Дополнительно был разработан еще один — тип шага. В данном исследовании при расчете текстурных дескрипторов лейкоцитов используются следующие типы шагов: классический, с использованием последовательностей и порядок модели МСП. Из четырех характеристик направленных матриц можно выбрать любое направление ($0^\circ, 45^\circ, 90^\circ$ и 135°), минимальное, максимальное или среднее значение. Усреднение дескрипторов четырех направлений позволяет добиться инвариантности характеристик к вращению, что является полезным при анализе клеток. При использовании второго из предложенных методов модификации МСВУЯ (на основе модели МСП) формируется только одна матрица. Таким образом, остается необходимость оптимизировать только один параметр — размер шага. Далее предлагается оригинальный подход к достаточно эффективному решению этого вопроса.

В данном случае имеется M классов объектов и N признаков. Необходимо вычислить несколько наборов признаков для разных значений шага, например, от 1 до 7. Задача состоит в выборе одного

из этих шагов, при котором достигается максимальное дифференцирование кластеров. Воспользуемся критерием Фишера для оценки расхождения двух кластеров по признаку k :

$$f_{ij}^k = \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2}, \forall ij, i > j, f_{ij}^k = f_{ji}^k.$$

Сформируем матрицу интегральной межклассовой оценки критерия Фишера:

$$F_{ij} = \sum_{k=1}^N f_{ij}^k,$$

где N — количество признаков.

Далее для каждого класса из этой матрицы находим минимальную оценку. Для этого анализируются строка и столбец матрицы, равные номеру класса:

$$F \min_d = \left\{ \min_{i=1}^M (F_{ij}, F_{ji}), j = 1 \dots M \right\}.$$

Таким образом, для каждого шага d построения текстурных дескрипторов будем иметь набор чисел размером, равным количеству классов M . Построим специальный график для оценки эффективности выбора шага формирования МСВУЯ (рис. 4 — см. на 3-й стороне обложки). Представленный график соответствует использованию МСВУЯ, модифицированных по второму методу. Каждый класс определяет свою ось координат. По осям откладываются полученные интегральные оценки Фишера для каждого шага. Соединенные значения оценок для одного шага образуют многоугольник. Значения отношения площади многоугольника к дисперсии соответствующего вектора $F \min_d$ представлены на рис. 4.

Предлагается выбрать шаг d построения текстурных дескрипторов, соответствующий максимальной приведенной площади полученного многоугольника. Эксперименты с данным методом показали высокую корреляцию его показаний с высшим средним процентом корректной классификации. В данном случае следует выбрать шаг, равный 4.

Таким образом, предложенный подход позволяет проводить автоматическую оптимизацию параметров формирования текстурных признаков с низкими вычислительными затратами. Это особенно актуально в условиях полученной размерности пространства признаков при достаточно большой тренировочной выборке.

Классификация

Так как проверка распределений значений признаков показала их близость к нормальному распределению (т. е. априори известен вид функции распределения), то оказывается возможным применение статистического классификатора, построенного на основе формулы Байеса.

Дискриминационная функция в случае нормального распределения примет вид

$$d_i(X) = -\frac{1}{2} \left(\ln (\Sigma_i) + (X - M_i)^T \Sigma_i^{-1} (X - M_i) \right),$$

где M_i — вектор математического ожидания; Σ_i — матрица ковариации.

Каждый кластер представляется гиперэллипсоидом, параметры которого описаны в выражении $(X - M_i)^T \Sigma_i^{-1} (X - M_i)$, называемом расстоянием Махалабониса. Главные оси гиперэллипсоида заданы собственными векторами Σ_i . Собственные числа Σ_i определяют длины этих осей. Центр гиперэллипсоида задан посредством M_i .

Большим преимуществом статистического классификатора является возможность вычисления вероятности ошибки классификации, с помощью которой можно оценить производительность выбранной дискриминационной функции. Вероятность ошибки в случае нормального распределения и симметричной функции потерь примет вид

$$P_e = \frac{1}{2} \Phi \left[-\frac{1}{2} \sqrt{r_{ij}} \right] + \frac{1}{2} \left(1 - \Phi \left[\frac{1}{2} \sqrt{r_{ij}} \right] \right) = \Phi \left[-\frac{1}{2} \sqrt{r_{ij}} \right].$$

Оценим качество кластеризации для выбранного в качестве оптимального метода построения текстурных дескрипторов.

В табл. 1 представлены значения межклассовых расстояний Махалабониса, полученных при применении второй модификации метода построения МСВУЯ и шага, равного 4. Минимальное расстояние — 122,48, что соответствует уровню ожидаемой максимальной ошибки вследствие перекрытия кластеров — 0,00000157 %. Это свидетельствует о достигнутом высоком качестве дифференцирования кластеров.

Полезный тест для проверки качества кластеров осуществляется путем классифицирования той же базы данных объектов, что была задействована при формировании параметров кластеров. Результаты такого теста составили 100 % корректной классификации для всех классов. Эти значения нельзя принимать в качестве основного результата, так как имеется прямая зависимость параметров классификатора от тестового массива. Однако на их основе можно сделать однозначный вывод об абсолютном разнесении кластеров в пространстве признаков, что подтверждает их (признаков) высокое дискриминационное качество. Также можно констатировать отсутствие явных кластерных патологий, таких как, например, возникновение кластера в другом кластере.

■ Таблица 1. Расстояния между классами (вторая модификация МСВУЯ, шаг 4)

	Моноцит	Лимфоцит	Нейтрофил	Базофил	Эозинофил
Моноцит	0	2500,21	481,10	347,87	175,24
Лимфоцит	3165,91	0	2443,70	352,55	11 033,34
Нейтрофил	713,53	1255,46	0	676,58	122,48
Базофил	1984,75	402,03	205,06	0	417,52
Эозинофил	215,42	1188,18	134,15	168,23	0

■ **Таблица 2.** Результаты вычисления T-статистики (вторая модификация МСВУЯ, шаг 4)

	Моноцит	Лимфоцит	Нейтрофил	Базофил	Эозинофил
Моноцит	0	57,23	35,03	41,82	34,38
Лимфоцит	57,23	0	70,79	51,45	54,19
Нейтрофил	35,03	70,79	0	33,71	22,84
Базофил	41,82	51,45	33,71	0	31,59
Эозинофил	34,38	54,19	22,84	31,59	0

Чтобы определить, действительно ли векторы M_i , определяющие центры кластеров, различны в статистическом смысле (т. е. принадлежат независимым генеральным совокупностям), применим обобщенную версию T-теста. Указанный тест — это многомерное расширение проверки с использованием статистики Стьюдента. Будем проверять гипотезу о равенстве пары векторов математических ожиданий (M_i, M_j), т. е. предположим, что выборки построения этих векторов взяты из одной генеральной совокупности. Уровень значимости при этом положим 1 %.

Для проверки гипотезы используется следующая статистика [18]:

$$t_{ij}^2 = \frac{n_i n_j (n_i + n_j - 2)}{n_i + n_j} (M_i - M_j)^T H_{ij}^{-1} (M_i - M_j),$$

где n_i — число объектов в классе i ; $H_{ij} = n_i \Sigma_i + n_j \Sigma_j$.

Данная статистика имеет $(n_i + n_j - 4 - 1)$ степеней свободы и распределена по следующему закону [18]:

$$T(x) = \frac{2\Gamma\left(\frac{n}{2}\right)}{(n-1)^{\frac{k}{2}} \Gamma\left(\frac{n-k}{2}\right) \Gamma\left(\frac{k}{2}\right)} \cdot \frac{x^{k-1}}{\left(1 + \frac{x^2}{n-1}\right)^{\frac{n}{2}}}, \quad x > 0,$$

где n — число степеней свободы статистики; k — порядок статистики.

Для выбранного уровня значимости была вычислена квантиль распределения — 7,766. Проверяемая гипотеза должна быть отвергнута при превышении статистикой этого значения.

Результаты вычисления статистики приведены в табл. 2. На их основании гипотеза о выборке данных из одной генеральной совокупности должна быть отвергнута. Таким образом, имеется пять статистически независимых популяций.

■ **Таблица 3.** Матрица неточностей классификации (вторая модификация МСВУЯ, шаг 4)

Клетка	Лимфоцит	Нейтрофил	Эозинофил	Базофил	Моноцит	Сумма
Лимфоцит	50	0	0	1	0	51
Нейтрофил	0	48	3	0	0	51
Эозинофил	0	0	46	0	0	46
Базофил	0	1	0	49	1	51
Моноцит	0	0	1	0	49	50
Неизвестный	0	1	0	0	0	1
Сумма	50	50	50	50	50	250
Процент корректного распознавания	100,0	97,9	92,0	98,0	98,0	
Средний процент корректного распознавания						97,2

Классический классификационный тест предусматривает разделение всего множества анализируемых объектов на две группы — тренировочную и тестовую (обычно равные). После этого определяются параметры классификатора (в данном случае описатели кластеров) с использованием тренировочной базы данных. Затем осуществляется проверка работоспособности классификатора на тестовой базе. Таким образом, достигается независимость настроек классификатора от объектов реальной работы. Данный подход высокоэффективен при наличии достаточно большого исходного массива анализируемых объектов.

При имеющемся размере базы данных применение классического метода не представляется возможным, так как разделение массива на две группы в таком случае резко снижает точность оценки результатов.

В данной ситуации целесообразно воспользоваться популярным среди исследователей методом leave-one-out, описанным, например, в [17]. Смысл его заключается в полном формировании всех кластеров, кроме одного. Из массива данных оставшегося класса извлекается один объект и формируется последний кластер. Классификатору предъявляется выбранный объект, при этом сохраняется независимость параметров классификатора от тестовых образцов. Далее процедура повторяется со всеми классами и объектами в них. Таким образом, в формировании результатов классификации последовательно принимают участие все объекты базы данных. Основной недостаток данного метода — вычислительная сложность, так как на каждом шаге необходимо рассчитывать параметры одного из кластеров заново.

Результаты

Результаты экспериментов с различными способами формирования МСВУЯ (классический, предложенный в [11], и обе модификации, предложенные в данной работе) представлены на рис. 5 (см. на 3-й стороне обложки). Гра-

фики показывают усредненные межклассовые проценты корректной классификации для трех методов. Видно, что максимальная производительность достигнута при помощи второй модификации классического метода построения МСВУЯ. Максимальный средний процент корректной классификации для этого метода составляет 97,2 % на шаге 4. Это согласуется с ожидаемым оптимальным шагом, выбранным по рис. 4.

Детально оценить результаты классификации можно с помощью матрицы неточностей (confusion matrix), представленной в табл. 3. Столбцы матрицы представляют собой результаты классификации клеток одного класса (можно отследить, с какими классами были «перепутаны» клетки данного класса в процессе классификации). Строки матрицы указывают на появление данного класса при классификации других классов. На-

пример, при классификации 50 нейтрофилов 48 раз они были распознаны верно, один нейтрофил был опознан как базофил и один раз зафиксирован отказ от классификации.

По табл. 3 видно, что наименьшая вероятность корректной классификации достигнута для эозинофила — 92,0. Следует отметить, что это достаточно редко встречающаяся клетка — около 2 % от общего количества лимфоцитов в пробе. Если положить количество лимфоцитов, найденных во время анализа, — 28 %, моноцитов — 9 %, нейтрофилов — 60 %, базофилов — 1 % и эозинофилов — 2 %, то по формуле полной вероятности получим процент корректной классификации на уровне 98,4 % (указанные цифры соответствуют средним процентным соотношениям клеток [1]).

В работе [19] описано исследование, призванное оценить уровень ошибки классификации, допускаемой квалифицированным экспертом. Для этого было задействовано девять экспертов, которым было предложено осуществить классификацию базы данных из 314 лейкоцитов. В состав экспертной группы входили пять профессиональных гематологов и четыре опытных лаборанта. Ошибка эксперта оценивалась относительно мажоритарного мнения группы. По отдельным классам максимальная экспертная ошибка составила 21,06 % (для эозинофила), а максимальная средняя (по всем экспертам) ошибка — 4,09 %. Усредненная по всем классам ошибка, характеризующая работу эксперта, в итоге составила 1,97 %.

Таким образом, полученные результаты автоматической классификации хорошо согласуются с результатами работы профессионального врача-гематолога.

Литература

1. **Абрамов М. Г.** Гематологический атлас. — М.: Медицина, 1979.
2. **www.cellatlas.com**
3. **Weszka J. S., Dyer C. R., Rosenfeld A. A.** Comparative Study of Texture measures for Terrain Classification // IEEE Transactions on Systems, Man and Cybernetics. — 1976. — N 6. — P. 269–285.
4. **Connors R. W., Harlow C. A.** A Theoretical Comparison of Texture Algorithms // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 1980. — N 2. — P. 204–222.
5. **Baf J. M. H., Kardan M., Spann N.** Texture feature performance for image segmentation // Pattern Recognition. — 1990. — Vol. 23. — N 3/4. — P. 291–309.
6. **Ohanian P. P., Dubes R. C.** Performance evaluation for four classes of textural features // Pattern Recognition. — 1992. — Vol. 25. — N 8. — P. 819–833.
7. **Augustejin M. F.** Performance evaluation of texture measures for ground cover identification in satellite images by means of neural-network classifier // IEEE Transactions on Geoscience and Remote Sensing. — 1995. — Vol. 33. — P. 615–625.
8. **Randen T., Husoy J. H.** Filtering for texture classification: a comparative study // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 1999. — Vol. 21. — N 4. — P. 291–310.
9. **Chen C. C., Chen P. C.** Filtering methods for texture discrimination // Pattern Recognition Letters. — 1999. — Vol. 20. — P. 783–790.
10. **Smith G.** Ohanian and Dubes Comparisons // www.cssip.uq.edu.au/staff/meastex/www/comparisons/ohanianDubesCmp.html
11. **Haralick R. M., Shanmugam K., Dinstein I.** Textural features for image classification // IEEE Transactions on Systems, Man and Cybernetics. — 1973. — Vol. SMC-3. — N 6. — P. 610–621.
12. **Pressman N. J.** Markovian analysis of cervical cell images // The Journal of Histochemistry and Cytochemistry. — 1976. — Vol. 24. — N 1. — P. 138–144.
13. **Parzen E.** Modern Probability Theory and Its Applications. — New York: John Wiley, 1969. — P. 136–147.
14. **Li S. Z.** Markov Random Field Modeling in Computer Vision. — Springer-Verlag, 1999.
15. **Besag J.** Spatial interaction and the statistical analysis of lattice systems (with discussions) // Journal of the Royal Statistical Society. — 1974. — N 36. — P. 192–236.
16. **Chellappa R., Chatterjee S.** Classification of textures using Gaussian Markov random fields // IEEE Transactions on Acoustics, Speech, and Signal Processing. — 1985. — Vol. ASSP-33. — N 4. — P. 959–963.
17. **Duda O. R., Hart P. E., Stork D. G.** Pattern classification (2nd edition). — Wiley-Interscience, 2000.
18. **Cramer H.** Mathematical Methods of Statistics. — Princeton Press, 1963.
19. **Bacus J. W., Gose E. E.** Leukocyte pattern recognition // IEEE Transactions on Systems, Man and Cybernetics. — 1972. — Vol. SMC-2. — N 4. — P. 513–526.