НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 1995, том 5, №1-2, с. 87 - 94

РАЗНАЯ ТЕМАТИКА

УДК 611.018.53.08

СТРУКТУРНО–ЛИНГВИСТИЧЕСКИЙ АНАЛИЗ МОРФОЛОГИИ ЛЕЙКОЦИТОВ

© 1995г. О.М. Горбенко

Институт аналитического приборостроения РАН, Санкт-Петербург

Поступила в редакцию 01.12.94

В работе описан алгоритм классификации лейкоцитов крови гранулоидной группы по стадиям зрелости (ранняя стадия созревания, палочкоядерный гранулоцит, сегментоядерный гранулоцит). Классификация проводится по морфологическим свойствам лейкоцитов. В основу предлагаемого алгоритма положен структурно– лингвистический метод распознавания образов. Приведены примеры обработки микроизображений лейкоцитов.

ВВЕДЕНИЕ

Задача проведения автоматизации проведения анализов крови представляется весьма актуальной в современных условиях, особенно вследствие увеличения числа заболеваний, связанных с изменением состава крови. Постановка этой задачи своевременна также потому, что высокопроизводительные ЭВМ, способные решать подобные задачи, стали широкодоступными. Автоматизация обработки анализов крови могла бы позволить осуществлять раннюю диагностику заболеваний крови путем проведения профилактических обследований больших групп населения, привлекая медицинских специалистов лишь для анализа патологических случаев, выявленных машиной. Эта задача находится в русле широко обсуждаемых проблем распознавания образов и машинного зрения.

Методы решения задачи распознавания распадаются на две группы: дискриминантный (числовой) подход и синтаксический (структурно-лингвистический) подход. При первом подходе объекты характеризуются наборами чисел (признаков), которые являются обычно результатами измерений, и распознавание образов (отнесение каждого объекта к некоторому классу) проводят путем разбиения пространства признаков на области. Развитие исследований по распознаванию клеток крови велось преимущественно в рамках числового подхода. Известны работы по выделению 5 видов лейкоцитов (нейтрофилы, зозинофилы, базофилы, лимфоциты, моноциты) на основании четырех признаков (размер ядра, цвет ядра, размер цитоплазмы, цвет цитоплазмы) [1], по делению лейкоцитов на 8 категорий (используется 8-размерный Гауссовский классификатор) [2]. В работе [3] вектор признаков строится с помощью морфологических операций над изображениями ядер лейкоцитов. В последнее время за рубежом разработаны и применяются автоматизированные системы классификации клеток крови, основанные на распознавании с применением дискриминантных методов [4,5].

Синтаксический подход был предложен К.С. Фу [6]. Этот метод разработан для решения задач распознавания образов, в которых важна информация, описывающая структуру объекта, взаимосвязь и расположение частей, из которых состоит объект. В биологических исследованиях синтаксический метод применялся для распознавания хромосом [6,7]. В задаче распознавания лейкоцитов анализируется структура клетки и форма ядра, поэтому в данной работе исследуется возможность применения структурно-лингвистического распознавания к этой задаче.

К достоинствам структурно-лингвистического распознавания можно отнести инвариантность относительно координатного распо-

ложения и ориентации объекта, отсутствие сложных вычислений, устойчивость к локальным изменениям.

При синтаксическом подходе считается, что образы состоят из соединенных различными способами подобразов, называемых непроизводными элементами. Правила композиции непроизводных элементов обычно задают при помощи грамматики языка описания образов. Процесс распознавания осуществляется после идентификации в объекте непроизводных элементов и составления описания объекта. Распознавание состоит в синтаксическом анализе "предложения", описывающего данный объект, т.е. процедуры, устанавливающей, является ли это предложение синтаксически правильным по отношению к заданной грамматике. Одновременно синтаксический анализ дает некоторое структурное описание объекта. Таким образом, систему синтаксического распознавания образов можно считать состоящей из трех основных частей: блок предобработки, блок описания или представления объекта и блок синтаксического анализа.

Блок предобработки осуществляет функции фильтрации, восстановления, улучшения, кодирования, аппроксимации объекта. После этого объект представляют в виде структуры языкового типа, например, цепочки. Этот процесс состоит из сегментации и выделения непроизводных элементов. Далее осуществляется синтаксический анализ и принимается решение о том, является ли представление объекта синтаксически правильным относительно заданной грамматики. Синтаксическое распознавание опирается на теорию формальных языков [8].

1 ВВОД ДАННЫХ И ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА

Объектами исследования являются окрашенные препараты анализов кроки, выполненные в виде мазков на стекле, наблюдаемые через оптический микроскоп. Ввод изображения в ЭВМ осуществляется с помощью черно-белой телекамеры на основе ПЗС-матрицы и системы обработки изображений SCPO (Польша). Для отладки алгоритма использовались также изображения типовых клеток из альбома [9]. Таким образом, на экране монитора получено полутоновое (256 градаций) черно-белое изображение (рис.1–3, вверху слева).

Исходными объектами являются полутоновые изображения клеток. Задачей этого этапа является получение контура объекта.

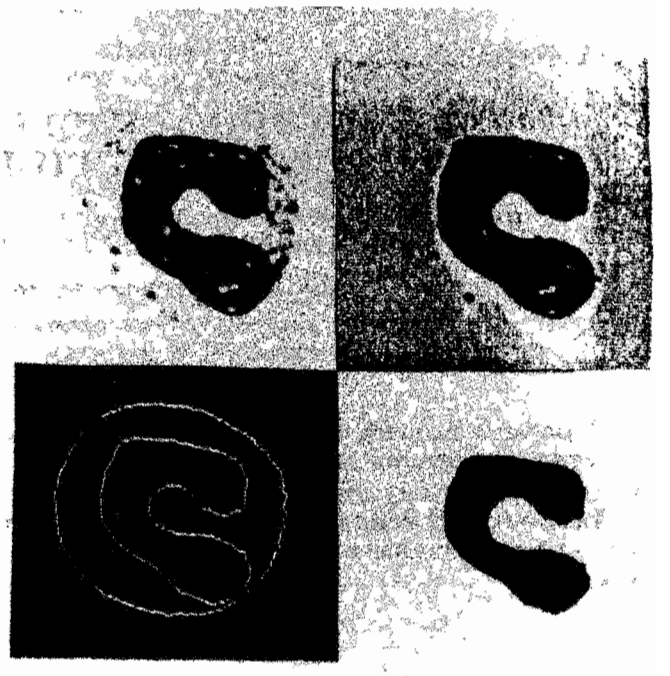

Рис.1. Исходное изображение и этапы предварительной обработки сегментоядерного гранулоцита из альбома [9].

Рис.2. Исходное изображение и этапы предварительной обработки палочкоядерного гранулоцита (анализ крови).

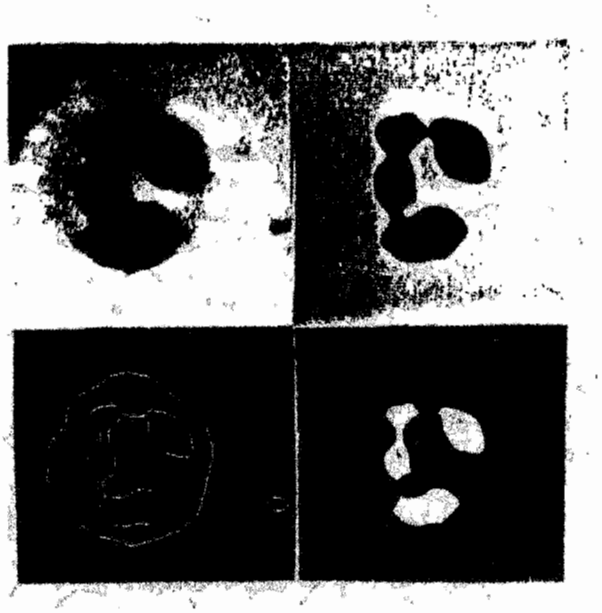

Рис.3. Исходное изображение и этапы предварительной обработки сегментоядерного гранулоцита (анализ крови).

Контур должен быть приведен к единичной толщине, т.е. быть 8–связным. По определению, 8–связным называется контур, в котором каждая точка в своей 8–окрестности имеет две и только две соседние точки.

Предварительная обработка проводится в среде системы обработки изображений SCPO, которая предоставляет широкий набор функций. В процессе предварительной обработки использовались следующие средства:

1) Увеличение контрастности (вычитание фона и умножение на константу);
2) Сглаживание с помощью свертки;
3) Пороговая обработка (для разделения ядра и цитоплазмы). Результат пороговой обработки представлен на рис.1–3 вверху, справа;
4) Морфологическая фильтрация. При необходимости бинарные изображения цитоплазмы и ядра, полученные в результате применения порогов, подвергаются морфологической фильтрации. Для этого применяются морфологические операции размыкания и замыкания со структурным элементом прямоугольной формы, соответствующим по площади величине устраняемых помех [10];
5) Скелетизация. В некоторых клетках ядра состоят из нескольких долей, соединенных нитями. Толщина нитей на изображении может быть неодинаковой, поэтому для приведения их к единичной толщине, как того требует алгоритм распознавания, применяется морфологическая скелетизация со структурным элементом размером 1 пиксел. Скелетизация проводится после инвертирования изображения ядра,так как после пороговой обработки интенсивность точек ядра равна нулю и циклически повторяется до тех пор, пока толщина всех нитей, соединяющих доли ядра, не становится равной 1;
6) Стирание "бахромы". После скелетизации на краях объектов изображения может появиться "бахрома", которую необходимо удалить. Для этого циклически выполняется операция стирания крайних точек;
7) Выделение контура. После выполнения необходимого набора из описанных выше операций (рис.1–3 внизу, справа) можно выделить контуры объектов (цитоплазмы и ядра), а затем получить контурное изображение клетки путем сложения контуров ядра и цитоплазмы. Чтобы выделить контур, необходимо определить граничные элементы [11]. Элемент с координатами (i, j) считается граничным только и только тогда, когда

$$a(i, j+1) + a(i, j-1) + \\ + a(i+1, J) + a(i-1, j) < 4 \quad .$$

Для получения контура изображения необходимо проанализировать все изображение по строкам, запоминая граничные элементы в массиве, после чего стереть изображение и вывести контур.

В том случае, если не удалось выделить область цитоплазмы при пороговой обработке, ее можно заменить произвольным замкнутым контуром или окружностью, чтобы возможно было применить распознающий алгоритм;
8) Стирание незамкнутых линий. Эта операция проводится с помощью того же алгоритма, что и стирание "бахромы", но уже в применении к контурным изображениям. Ее необходимо провести, если на изображении остались части соседних с исследуемым объектов, а также для очищения контура исследуемой

клетки;

9) Утоньшение контура. Применяется тот же алгоритм, что и при скелетизации бинарного изображения. При этом стираются все лишние точки, и в результате получается контур единичной толщины (см.рис.1–3 внизу, слева).

Таким образом, проведена предварительная обработка и сегментация изображения. Теперь необходимо провести выделение непроизводных элементов, составление из них цепочки и синтаксический анализ.

2 СТРУКТУРНО-ЛИНГВИСТИЧЕСКОЕ РАСПОЗНАВАНИЕ ОБРАЗОВ В ПРИМЕНЕНИИ К НЕКОТОРЫМ ТИПАМ ЛЕЙКОЦИТОВ

Постановка задачи. Предлагается алгоритм распознавания нескольких типов клеток крови миэлоидного ряда по контурному изображению. Таким образом, алгоритм не учитывает цветовые и текстурные особенности цитоплазмы и ядра, а осуществляет классификацию лишь по форме ядра клетки.

Ядро миэлоидной клетки, или гранулоцита, меняется в процессе созревания клетки, и его форма служит характеристикой стадии ее развития. У незрелых клеток форма ядра округлая, затем, по мере созревания, ядро становится все более вырезанным, и когда размер выреза становится примерно в половину гипотетического круглого ядра, и противоположные края ядра становятся параллельными на заметном расстоянии, гранулоцит называют палочкоядерным (band) [9]. По мере дальнейшего созревания вырезанность ядра еще более усиливается, и ядро разделяется на доли, соединенные тонкими волокнами. Такие клетки называются сегментоядерными (segmented) [9]. Переход между различными стадиями развития клеток постепенный, и они часто трудноразличимы. В случае сомнений правило таково: поместить клетку в более зрелую категорию [9].

В предложенном алгоритме классификации клеток одиночный контур, не имеющий внутренних, считается безъядерной структурой, пара контуров, один из которых вложен в другой — клеткой с ядром. Тип клетки зависит от формы ядра.

Предлагаемый алгоритм может распознавать следующие типы объектов: безъядерная структура, незрелая форма клетки, палочкоядерный гранулоцит, сегментоядерный гранулоцит. При необходимости его можно настроить на распознавание других типов клеток, либо расширить набор классифицируемых клеток.

Описание исходных данных. Исходные данные для алгоритма распознавания могут быть получены в результате предварительной обработки полутоновых изображений, процедуры которой описаны выше.

Таким образом, в левом нижнем углу экрана монитора, в окне размером 210 на 210 пикселов, должно располагаться контурное изображение исследуемого объекта, и контур должен быть единичной толщины (8-связный).

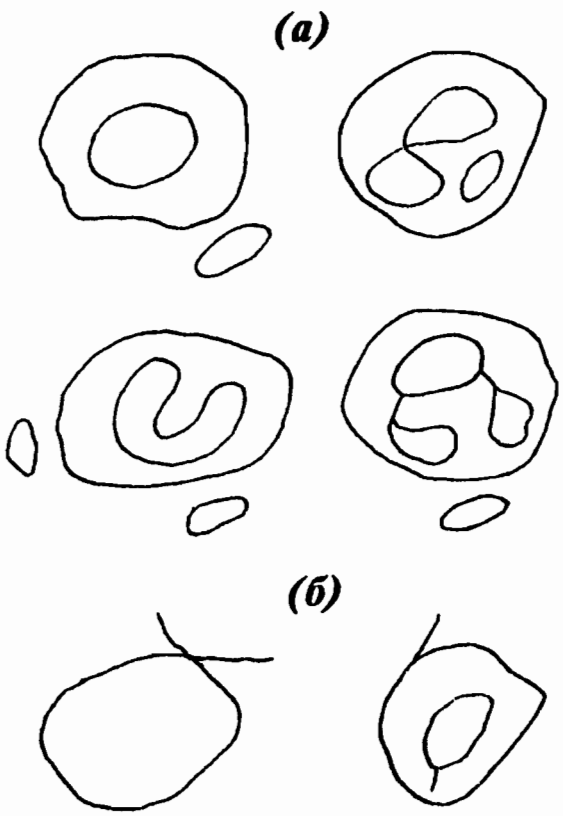

Рис.4. Примеры допустимых (a) и недопустимых (b) алгоритмом синтаксического распознавания изображений.

Допустимыми являются простые одиночные контуры, вложенные контуры, контуры, имеющие точки самопересечений, и контуры, соединенные линиями. Не допускается наличие незамкнутых линий. Примеры изображений приведены на рис.4. Чтобы из приве-

денных на рис.4,б изображений получить допустимые, надо использовать функцию, описанную в пункте 8 части 1.

Описание работы алгоритма. Блок-схема алгоритма синтаксического распознавания представлена на рис.5.

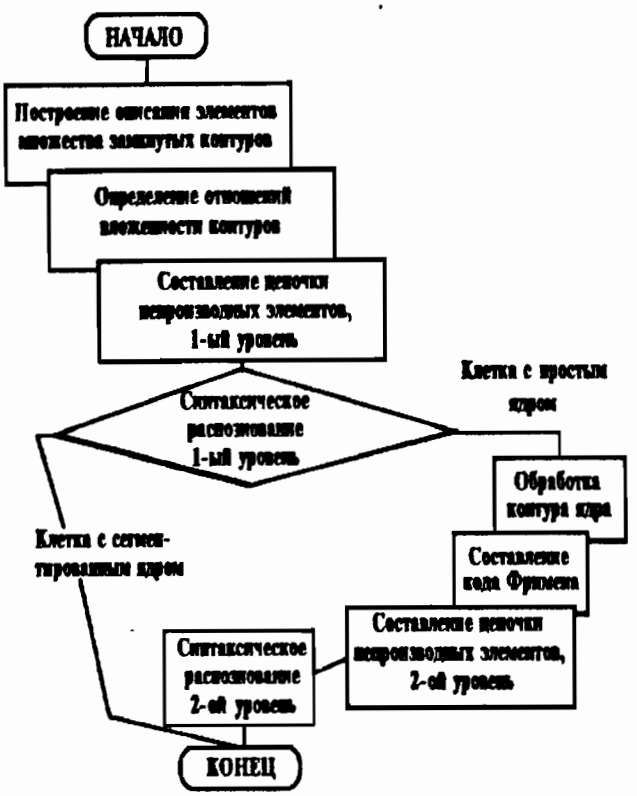

Рис.5. Блк-схема алгоритма синтаксического распознавания изображений лейкоцитов.

Сначала осуществляется сканирование заданного окна на экране с целью нахождения первой точки контура. Затем производится обход контура и стирание его. Если его размеры больше некоторого минимального размера, координаты точек контура запоминаются в массиве структур CONTUR. Шаблон структуры CONTUR предусматривает следующую информацию о кнтуре: координаты точек, количество точек в контуре, количество замкнутых соединенных составляющих (долей), информация о вложенности. Таким образом обрабатываются все контуры, имеющиеся на изображении, и в результате создается массив структур, содержащий информацию о них. Отношения вложенности фиксируются в массиве структур CONTUR.

После создания структур данных можно провести структурно-лингвистическое распознавание. Предлагается иерархическое применение этого метода [11]. На первом уровне описывается структура клетки в целом, а на втором – структура ядра, если это необходимо для классификации. Введем определения некоторых ключевых понятий теории формальных языков [8].

Алфавит – любое конечное множество символов. Предложение (цепочка, слово) в некотором алфавите – произвольная цепочка конечной длины, состоящая из символов этого алфавита. Предложение, не содержащее ни одного символа, называется пустым предложением и обозначается s_0. Язык – произвольное множество предложений в некотором алфавите. Грамматикой называют четверку

$$G = (V_N, V_T, P, S) \quad ,$$

где V_N – множество нетерминальных символов; V_T – множество терминальных символов; P – множество грамматических правил, или правил подстановки; S – начальный или корневой символ.

Предполаается, что S принадлежит множеству V_N и что V_N и V_T – непересекающиеся множества. Алфавит V является объединением алфавитов V_N и V_T . Между множествами терминальных символов и непроизводных элементов устанавливается взаимооднозначное соответствие.

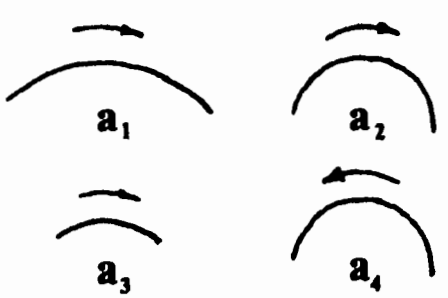

Рис.6. Множество непроизводных элементов для грамматик II уровня.

В качестве грамматик I уровня можно предложить G_1 и G_2 . Введем обозначение: X@Y означает, что объект Y находится внутри

объекта X.

$G_1 =$ $\{P_1, V_{N1}, V_{T1}, S_1\}$

$S_1 =$ < сегментоядерный гранулоцит >

$V_{N1} =$ {< оболочка >, < ядро >, S}

$V_{T1} =$ {< множество соединенных контуров >, < замкнутый контур >}

$P_1 :$ $S \rightarrow$ < оболочка > @ < ядро > < оболочка >$\rightarrow$ < замкнутый контур > < ядро >$\rightarrow$ < множество соединенных контуров >

$G_2 =$ $\{P_2, V_{N2}, V_{T2}, S_2\}$

$S_2 =$ < клетка с простым ядром >

$V_{N2} =$ {< оболочка >, < ядро >}

$V_{T2} =$ {< замкнутый контур >}

$P_2 :$ $S_2 \rightarrow$ < оболочка > @ < ядро > < оболочка >$\rightarrow$ < замкнутый контур > < ядро >$\rightarrow$ < замкнутый контур >

Если изображение, находящееся на экране, допустимо грамматикой G_2 , то необходимо провести второй этап распознавания. На втором уровне предлагается набор непроизводных элементов, представляющих из себя дуги окружностей различных длин и радиусов (рис.6). При настройке алгоритма на другие объекты множество непроизводных элементов легко изменить или дополнить.

Рассмотрим грамматики для каждого типа ядер, которые требуется распознать. Для описания ядра, встречающегося в незрелых формах клеток и имеющего выпуклую округлую форму, можно предложить следующую грамматику:

$$G_1 = (V_{N1}, V_{T1}, P_1, S_1) \quad ,$$

где $V_{N1} = \{S_1\}$, $V_{T1} = \{a_1, a_2, a_3\}$, а множество P состоит из следующих правил:

$$S_1 \longrightarrow a_1 S_1, \quad S_1 \longrightarrow a_2 S_1,$$
$$S_1 \longrightarrow a_3 S_1, \quad S_1 \longrightarrow a_1,$$
$$S_2 \longrightarrow a_2, \quad S_1 \longrightarrow a_3.$$

Пример цепочек, порождаемых грамматикой G_1 : $a_2a_1a_1, a_1a_3a_2a_1a_1, a_3a_1a_2$. Таким образом, синтаксически правильными по отношению к грамматике G_1 являются цепочки, состоящие из любого количества элементов a_1, a_2, a_3, встречающихся в любой последовательности.

Для описания формы ядра палочкоядерного гранулоцита предлагается грамматика

$$G = (V_{N2}, V_{T2}, P_2, S_2) \quad ,$$

где $V_{N2} = \{S_2, A, B, C, s_0\}$. Здесь s_0 – пустая цепочка, $V_{T2} = \{a_1, a_2, a_3, a_4\}$. Следующим этапом лингвистического распознавания является грамматический разбор цепочек. Приведем примеры восходящего разбора, при котором из цепочки терминальных символов путем перебора правил из множеств P_1 или P_2 осуществляется попытка достичь соответствующего корневого символа (рис.7). На рис.7,а показана попытка применить правила грамматики G_2 к цепочке $a_2a_1a_1a_3a_1$. При этом начальный символ получить не удалось, поэтому данная цепочка не принадлежит этой грамматике. При применении правил грамматики G_1 к цепочке рис.7,б получен корневой символ этой грамматики, значит, цепочка принадлежит этой грамматике. Цепочка на рис.7,в, аналогично, принадлежит грамматике G_2 .

На практике грамматический разбор осуществляется только если для распознавания требуется полное описание объекта. В противном случае для увеличения эффективности можно использовать более простые подходы (сравнение цепочки с эталоном, проверка наличия определенных подобразов и др.) [6].

В данном применении алгоритма признаки объекта можно определить непосредственно по составленной цепочке непроизводных элементов, так как один тип клеток характеризуется округлым выпуклым ядром (незрелая форма клетки), а другой – сильно вырезанным ядром (палочкоядерный гранулоцит). Поэтому при обнаружении в цепочке терминального символа, соответствующего вогнутой дуге, просмотр цепочки можно прекратить и сделать вывод, что клетка — палочкоядерный гранулоцит.

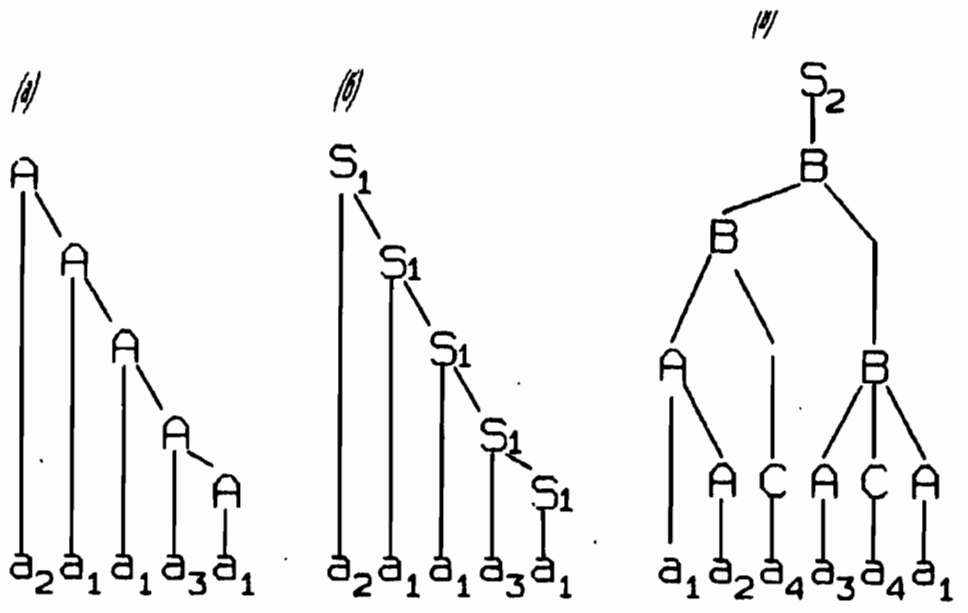

Рис.7. Примеры восходящего грамматического разбора цепочек, описывающих форму ядра лейкоцита.

При расширении множества типов распознаваемых объектов и усложнении их структуры грамматический разбор цепочки терминальных символов необходимо осуществлять полностью.

ВЫВОДЫ

Таким образом, предложен алгоритм, реализующий синтаксическое распознавание в применении к изображениям лейкоцитов крови группы гранулоцитов. Он осуществляет их классификацию по стадиям зрелости. Использованный метод позволяет легко настроить алгоритм на распознавание и других типов клеток.

Для усовершенствования предложенного метода распознавания изображений лейкоцитов и введения количественных оценок объектов можно воспользоваться механизмом аттрибутивных грамматик [11], при котором синтаксическим правилам грамматик могут быть поставлены в соответствие семантические правила, задающие количественные соотношения подобразов.

СПИСОК ЛИТЕРАТУРЫ

1. *Young T.* The Classification of White Blood Cells // IEEE Trans. on Biomed. Engineering. 1972. Vol.19, №4. P.291–298.

2. *Bacus W.* Leukocyte Pattern Recognition // IEEE Trans. on System. Man and Sybern. 1972. Vol.2. P.513–526.

3. *Киричук Н.А., Косых В.П., Петунин А.Н.* Количественный анализ миэлоидных клеток человека // Автометрия. 1989. №2. С.34.

4. *Zajicek G., Shoat M.* On the Classification of Nucleated Red Blood Cells // Comput.& Biomed. Res. 1983. Vol.16, №6. P.553–562.

5. *Parthenis K.* An Automatic Computer vision system for blood analysys // Microprocess. Microprogr (Netherlands). 1990. Vol.28, №1–5. P.243–246.

6. *Фу К.* Структурные методы в распознавании образов: Пер.с англ. М., 1977. 317 с.

7. *Tsai Wen-Hsiang* Attributed Grammar - a Tool for Combining Syntactic and Statistical Approaches to Pattern Recognition // IEEE Trans. on System, Man and Cybern. 1980. Vol.10, №12.

8. *Ту Дж., Гонсалес Р.* Принципы распознавания образов : Пер. с англ. М., 1978. 411 с.

9. *Diggs L.W., M.D., Sturm Dorothy* The Morphology of Blood Cells // Memphis, Tennesse, 1954.

10. *Марагос П., Шафер Р.У.* Морфологические системы для многомерной обработки сигналов // ТИИЭР. 1990. Т.78, №4. С.109.

11. *Семенков О.И., Абламейко С.В., Берейшик В.И., Старовойтов В.В.* Обработка и отображение информации в растровых графических системах. М., 1989. С.180.

12. *Pavlidis T.* Syntactic Recognition of Handwritten Numerals // IEEE Trans. on System, Man and Cybern. 1977. Vol.7, №7. P.537.