ОБРАБОТКА ИЗОБРАЖЕНИЙ, СИГНАЛОВ И РЕЧИ

УДК 004.93; 519.68

В.А. Ковалев, А.А. Дмитрук, И.В. Сафонов

МЕТОД ПОИСКА СВЯЗЕЙ МЕЖДУ МОРФОЛОГИЧЕСКИМИ СТРУКТУРАМИ ГИСТОЛОГИЧЕСКИХ ИЗОБРАЖЕНИЙ И ПОКАЗАТЕЛЯМИ СОСТОЯНИЯ ОНКОЛОГИЧЕСКИХ БОЛЬНЫХ

Предлагается метод поиска и визуализации структур, связанных с показателями состояния онкологических больных. В основе метода лежит многоступенчатая процедура, включающая подсчет признаков изображения, извлечение главных компонент, корреляцию главных компонент с известными свойствами объекта и проекцию найденных зависимостей на исходные гистологические снимки с целью выделения информативных структурных образований. Находятся зависимости между клиническими показателями и морфологическими структурами на соответствующих изображениях.

Введение

Визуальный осмотр гистологических изображений является «золотым стандартом» при постановке окончательного диагноза [1, 2], определении стадии и лечения в случае многих видов рака. При этом в общей дисциплине анализа биомедицинских изображений проблеме анализа гистологических изображений до сих пор уделяется недостаточно внимания [3–5]. Основная причина этого заключается в заметной морфологической сложности, которая присуща микроскопным снимкам тканей живых организмов [3, 4]. В этой связи авторами рассмотрена задача нахождения соответствий между диагностическими показателями онкологических больных и морфологическими структурами на гистологических изображениях ткани яичника.

Выбор типа анализируемых изображений обусловлен тем, что рак яичника – разрушительное заболевание, признанное одной из главных причин женских гинекологических смертей в мире [6]. Начальные признаки злокачественной опухоли в виде небольшого и безболезненного увеличения живота остаются без внимания, и первое обращение женщин в больницы происходит обычно на поздних стадиях заболевания. Поэтому пятилетняя выживаемость в этом случае не превышает 20 % [6, 7].

Наряду с изображениями образцов ткани яичника, окрашенных гематоксилин-эозином (рутинная окраска), были рассмотрены изображения, окрашенные маркером D2-40. Этот маркер выделяет лимфатические сосуды в тканях яичника (если точнее, то он является селективным маркером лимфатического эндотелия в нормальных тканях и патологических сосудистых).

Излагаемый метод разработан в рамках большого проекта, направленного на изучение ангиогенеза злокачественных опухолей в ткани яичника [8]. Ангиогенез как развитие новой системы кровеносных сосудов – это важный фактор существенного увеличения опухоли и метастазирования [7]. Без ангиогенеза размер опухоли естественным образом ограничивается 1–2 мм, поскольку для дальнейшего ее роста необходим кислород и питание [9]. Надежды на лечение рака связаны с поиском способов торможения процесса ангиогенеза. Отсюда следует, что обнаружение связей между структурой опухоли, характеристиками ее роста и состоянием пациента является первостепенной задачей для онкологии [6, 7].

В клинической практике предусмотрен сбор информации в базу данных пациентов, содержащую изображения различных модальностей и показатели онкологических больных. К таким показателям относятся, например, социальный статус пациента, результаты обследований и лабораторных тестов до и после операции, история болезни и т. п. Техническая проблема состоит в поиске статистически значимых зависимостей между морфологическими структурами на изображениях, представленными в форме количественных характеристик, и записями в таблице показателей пациентов. Такие корреляции могут быть найдены непосредст-

венно с помощью известных методов многомерного статистического анализа. Однако эта техника применима только для априорного анализа, когда существуют предположения о возможных взаимосвязях между структурами на изображении и состоянием пациента и статистический анализ применяется с целью подтвердить или опровергнуть эти гипотезы. Данный подход использовался авторами на этапе предварительного исследования. Осуществлялся поиск соответствий между состоянием пациента и структурами, отмеченными маркером D2-40. Для этого лимфатическая сеть была сегментирована и охарактеризована пятью количественными признаками. Далее был проведен поиск корреляций признаков с данными пациентов. Полученные частные и достаточно скромные выводы не оправдали затраченное время и ресурсы на решение данной задачи [8]. Поэтому рассмотрен альтернативный исследовательский подход, который позволяет, во-первых, обнаружить множество объективно существующих связей и, во вторых, интерпретировать найденные результаты с медицинской точки зрения. Таким образом, работа выполнена в области интеллектуального анализа изображений (image mining). Подобно интеллектуальному анализу данных (data mining) анализ изображений может быть интерпретирован в этом смысле как процесс обнаружения скрытых и неочевидных структур [10, 11] и извлечения из изображений связей и данных, не хранящихся в базе в явном виде [12–14]. В силу морфологической и структурной сложности гистологических изображений процесс их анализа сложно автоматизировать. Поэтому всесторонний интеллектуальный анализ данных является многообещающим для решения поставленных задач.

Для получения ожидаемых результатов разрабатываемый метод поиска связей между морфологическими структурами гистологических изображений и показателями состояния онкологических больных должен удовлетворять следующим требованиям:

1) дескрипторы изображений должны быть в достаточной мере информативными для адекватного представления широкого диапазона морфологических признаков изображений, как цветных, так и представленных оттенками серого;

2) количественные признаки, вычисленные с помощью дескрипторов и коррелированные с данными пациентов, должны допускать несложное отображение отобранных корреляций на оригинальные изображения для визуализации ключевых морфологических структур;

3) число признаков для описания одного изображения не должно быть намного больше общего количества пациентов. Это требование позволит избежать появления случайных корреляций, которые неизбежны для очень больших наборов данных.

Целью работы является разработка метода поиска связей, удовлетворяющего указанным выше условиям. Использование метода демонстрируется на анализе базы данных пациенток, страдающих раком яичника.

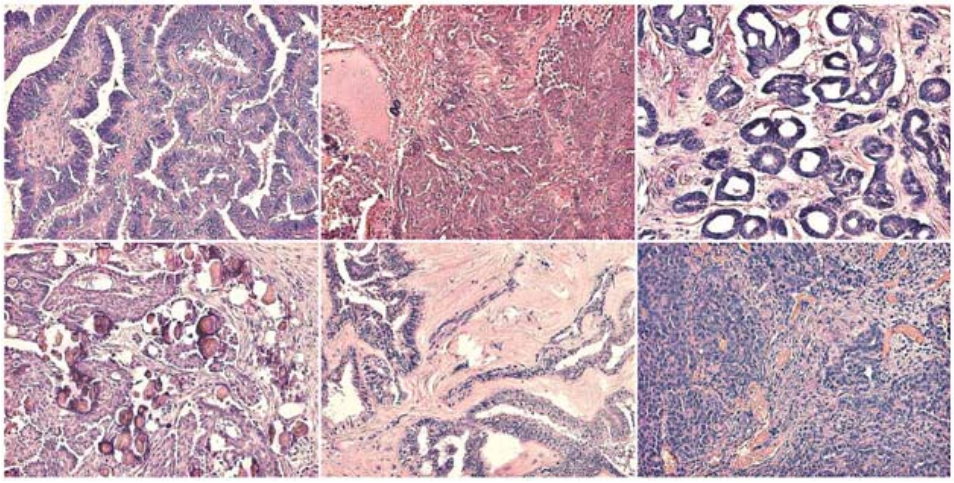
1. Материалы и методы

База данных состояла из таблицы с показателями состояния больных и гистологических изображений опухолевой ткани яичника по 68 пациентам онкологического диспансера (женщины, средний возраст 59,8 лет, среднеквадратическое отклонение возраста 11,2). Набор гистологических изображений представлял собой 952 полноцветных снимка размером 2048×1536 пикселов, полученных с помощью микроскопа Leica DMD108 при 200-кратном увеличении. В состав набора включены 272 изображения с рутинной прокраской (гематоксилином и эозином), по четыре изображения на каждого пациента, и 680 изображений ткани яичника, обработанной маркером D2-40, по 10 изображений на каждого пациента. Примеры всех типов изображений показаны на рис. 1.

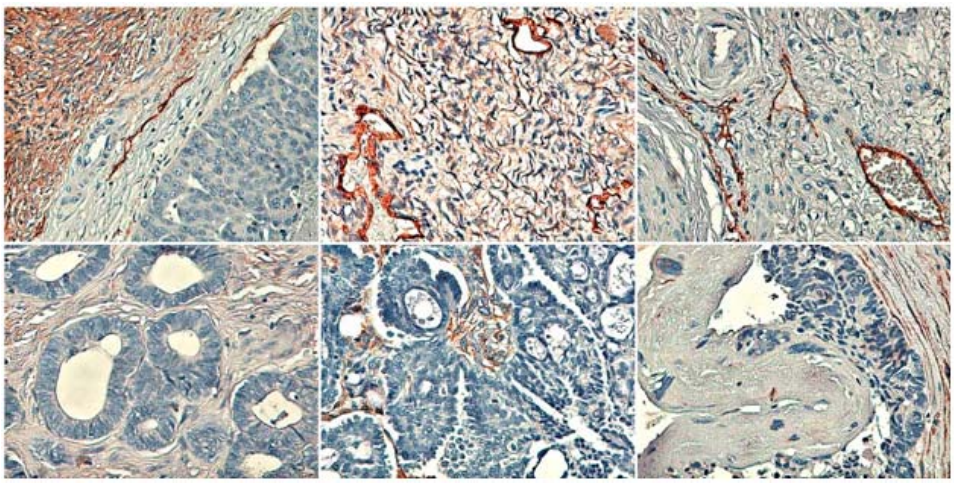
Таблица с показателями больных содержала около 80 различных признаков (TNM-классификацию опухоли, количество курсов химиотерапии, дифференциацию опухоли, детали хирургических операций и т. п.).

Из-за упоминаемых выше особенностей гистологических изображений для их анализа наиболее часто используются дескрипторы не отдельных областей или сегментов, а изображения в целом. Хорошо зарекомендовали себя цветные матрицы совместной встречаемости, предложенные независимо в работах [15, 16]. Однако принимая во внимание первое требование из введения настоящей статьи, подход с совместной встречаемостью пикселов был модернизи-

рован до четырехмерного случая, а именно: для определения соседства использовались тройки пикселов, лежащих в вершинах равносторонних треугольников с заданной длиной стороны. Такое расширение базового подхода технически не сводится только к добавлению дополнительного измерения к матрице совместной встречаемости. Отличия гораздо глубже и связаны с проблемой описания разнообразных типов текстур с помощью статистик первого порядка (яркостей пикселов или значений цветов), второго (градиентов) и более высоких порядков. Эта проблема была тщательно исследована Б. Джулисом [17]. Ученым было экспериментально доказано [18], что паттерны активности мозга значительно различаются при наблюдении текстур с низкими и высокими пространственными корреляциями. При этом важно различать статистики высокого порядка в пространственном и в яркостном [19] смыслах.



а)



б)

Рис. 1. Примеры изображений: *а)* рутинная прокраска; *б)* прокраска D2-40

Формализуем используемые дескрипторы следующим образом. Пусть $F_G = \{I(x,y)\} = \{I(i)\} = \{I(j)\} = \{I(k)\}$ – изображение в оттенках серого размером $M \times N$ пикселов. Положим также, что все пикселы изображения пронумерованы с помощью индексов i, j и k, где $i = \overline{1, MN}$, $j = \overline{1, MN}$, $k = \overline{1, MN}$ и их уровни интенсивности равны $I(i)$, $I(j)$, $I(k)$ соответственно. Индексы естественным образом определяются координатами соответствующих пикселов: $i = (x_i, y_i)$, $j = (x_j, y_j)$, $k = (x_k, y_k)$. Тогда для серого изображения четырехмерная матрица совместной встречаемости типа *IIID*, которая задается тройками пикселов (i, j, k), распо-

ложенных в вершинах равносторонних треугольников с длинами сторон из множества D, может быть определена следующим образом:

$$W_{\Delta D} = \left\| I(i), I(j), I(k), d \right\|; \tag{1}$$

$$d(i,j) = d(i,k) = d(j,k),\ d \in D; \tag{2}$$

$$i < j,\ i < k, \forall i : y_j \geq y_i, y_k < y_i. \tag{3}$$

Неравенства (3) заключают в себе требование перебора всех возможных треугольников без повторений. Система уравнений (1)–(3) описывает алгоритм покрытия всего изображения равносторонними треугольниками. Как следует из этой системы, процедура покрытия состоит из последовательных смещений в каждую точку i изображения и построения вокруг этой точки равносторонних треугольников с заданной длиной стороны. При этом две другие вершины треугольника не должны оказаться выше базовой вершины i.

Когда рассматриваются не яркости пикселов, а их цвета, исходное цветовое пространство редуцируется и соответствующая цветная матрица совместной встречаемости типа *CCCD* определяется по той же схеме с использованием индексов цветов вместо уровней яркости. После вычисления матрицы встречаемости обычно считаются признаки Харалика, и далее они используются на этапах описания, кластеризации и т. д. Однако данная процедура в рассматриваемом случае неприменима, поскольку признаки Харалика не могут быть отображены обратно на исходные изображения, что является необходимым условием и декларировано во втором требовании к используемым признакам в этой статье. Сами элементы матрицы совместной встречаемости могут быть отображены на исходное изображение [20], но их слишком много (более 7000), чтобы удовлетворить третьему требованию, предъявляемому к дескрипторам. Решение этого противоречия приводит к использованию метода главных компонент (Principal component analysis – PCA). С помощью PCA из элементов матрицы совместной встречаемости извлекается небольшое число некоррелированных между собой признаков.

Разработанный метод заключается в подсчете четырехмерных матриц совместной встречаемости, извлечении главных компонент, корреляции полученных компонент с данными пациента, выборе среди этих данных высококоррелированных значений и обратном их проецировании на матрицу совместной встречаемости. Последним шагом метода является использование выделенных таким образом элементов матрицы совместной встречаемости для визуализации соответствующих структур на гистологических изображениях.

Поскольку главные компоненты некоррелированы по определению, поиск значимых связей выполнялся с помощью простых однофакторных статистических методов, таких как корреляционный анализ главных компонент и показателей пациентов, а также t-тест Стьюдента. Последний применялся в случаях, когда значения показателей состояния пациента, например таких как «наличие удаленных метастаз» (присутствуют/отсутствуют), «степень дифференциации раковой опухоли» (высокая/низкая) и т. д., были представлены в номинальной шкале и разбивали естественным образом пациентов на группы. В указанных случаях с помощью t-теста проводилось сравнение значимости отличия средних значений полученных главных компонент по выделенным подгруппам пациентов.

2. Результаты исследования

Оригинальные изображения в формате RGB были конвертированы в цветовое пространство Lab, количество исходных цветов было уменьшено до 24 наиболее представительных с помощью алгоритма median cut [21]. Этот алгоритм используется в области обработки изображений для эффективного редуцирования цветового пространства. Вначале трехмерные расширенные цветовые матрицы типа *CCC* с фиксированным межпиксельным расстоянием d содержали $24^3 = 13\,824$ элемента. С учетом дальнейшего приведения матрицы к нижнему треугольному виду количество используемых элементов матрицы для одного межпиксельного расстояния сократилось до $N = 2600$. Поскольку одновременно использовались три типа равно-

сторонних треугольников с длинами сторон $D = \{1, 2, 3\}$, окончательное количество анализируемых элементов для каждого изображения равнялось $2600 \times 3 = 7800$. Далее была составлена сборная таблица признаков, состоящая из 7800 столбцов (по количеству признаков на каждом изображении) и 68 строк (по количеству пациентов). Каждая строка этой таблицы представляла собой векторизованную матрицу совместной встречаемости изображения. После применения метода главных компонент для дальнейшего анализа были выбраны 27 компонент для рутинных изображений и 38 – для изображений с маркером D2-40, покрывающих 95 % вариабельности исходных признаков. Эти результаты позволяют предположить, что структурная вариабельность изображений с маркером D2-40 существенно выше по сравнению с рутинными.

После корреляции с клиническими данными 27 главных компонент признаков рутинных изображений выявлено 43 корреляции с уровнем значимости $p < 0{,}01$. Та же процедура, примененная к 38 главным компонентам признаков другого типа изображений, выявила 47 значимых связей и записей в таблице состояния больных. При детальном изучении значимых корреляций выяснилось, что не во всех случаях найденные связи между определенными гистологическими структурами и клиническими показателями могут быть легко интерпретированы, несмотря на предполагаемый интерес с научной и практической точек зрения. Например, при анализе корреляций между главными компонентами признаков рутинных изображений и данными состояния больных выглядят многообещающими следующие показатели: развитие удаленных метастаз ($p<0{,}001$), степень дифференциации раковой опухоли ($p < 0{,}007$), количество выкидышей ($p<0{,}0001$) и количество проведенных курсов химиотерапии ($p < 0{,}000002$, $r = -0{,}543$). Визуализация некоторых структур приведена в верхней строчке рис. 2.

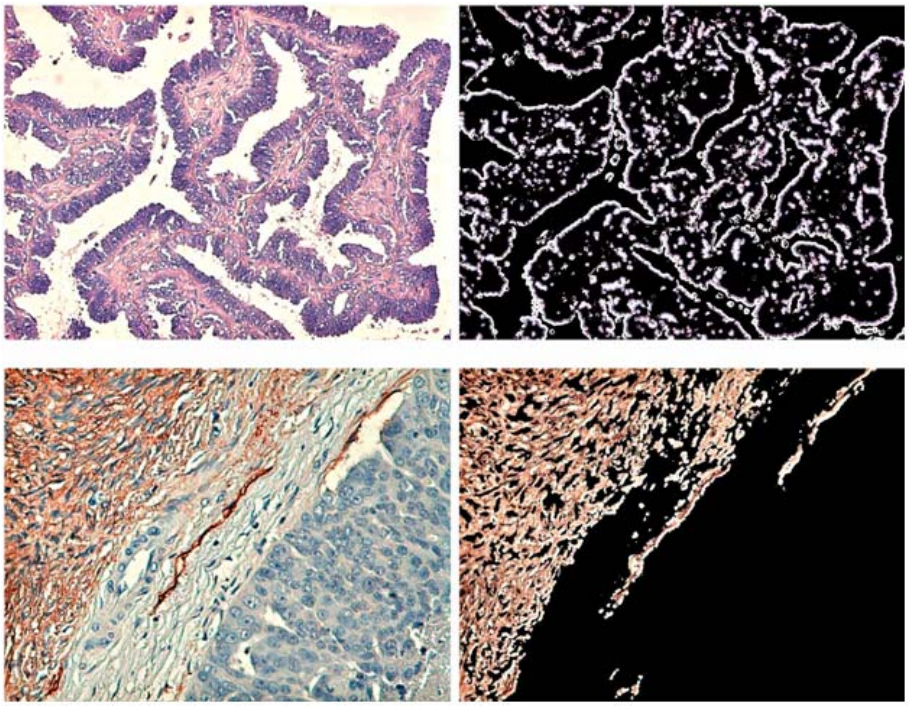


Рис. 2. Исходные изображения (левая колонка) и найденные ключевые структуры (правая колонка) для рутинных (верхняя строка) и D2-40 (нижняя строка) изображений

Обратная зависимость длины границы с количеством обследований может быть объяснена фактом, что у более «молодых» опухолей, которые поддаются химическому лечению, структура ткани более развита по сравнению со «старыми» опухолями, которые сразу удаляются. Изображения найденных связей для тканей, обработанных эндотелиальным маркером D2-40, демонстрируют похожие признаки. На нижней строчке рис. 2 показана одна из таких зависимостей. Доля прокрашенных маркером D2-40 структур сильно коррелирует с уровнем дифференциации опухоли ($p < 0{,}009$), временем выживаемости пациента ($p < 0{,}010$) и наличием рецидива ($p < 0{,}017$).

Также вычислялись совместные матрицы встречаемости типа *IIID* на конвертированных в оттенки серого изображениях. Несмотря на несколько найденных многообещающих корреляций для этого типа дескрипторов, в целом полученные данные противоречивы. В частности, при обратном отображении элементов матрицы *IIID* на исходное серое изображение одновременно выделялись структуры с различным биологическим смыслом. Причина этого видится в неизбежном при таком подходе редуцировании диапазона различных цветов в одно значение уровня яркости серого цвета.

Заключение

Представленный метод может рассматриваться как многообещающий инструмент для автоматической идентификации и визуализации структур на гистологических изображениях, релевантных данным о состоянии пациентов.

Поскольку в методе не предусмотрено внутренних средств для интерпретации найденных связей, необходим экспертный анализ результатов.

Дальнейшее усовершенствование метода предполагает развитие автоматической процедуры отбора элементов матрицы для отображения на исходные изображения после расчета главных компонент.

Работа выполнена в рамках проекта МНТЦ В-1682.

Авторы выражают благодарность сотрудникам Минского городского клинического онкологического диспансера М.В. Фридману и С.Е. Шелкович за предоставленные данные и консультации по медицинским вопросам.

Список литературы

1. Schwab, M. Encyclopedia of Cancer / M. Schwab. – N.Y. : Academic Press, 2009. – 3235 p.

2. Hayat, M. Methods of Cancer Diagnosis, Therapy and Prognosis. In 6 vol. / M. Hayat. – Springer, 2009–2010.

3. Wootton, R. Image Analysis in Histology: Conventional and Confocal Microscopy / R. Wootton, D. Springall, J. Polak. – Cambridge : Cambridge University Press, 1995. – 425 p.

4. Histopathological image analysis : A review / M.N. Gurcan [el al.] // IEEE Reviews in Biomedical Engineering. – 2009. – Vol. 2. – P.147–171.

5. Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading / O. Sertel [et al.] // Journal of Signal Processing Systems. – 2009. – Vol. 55, № 1. – P.169–183.

6. Stack, M.S. Ovarian Cancer (Cancer Treatment and Research) / M.S. Stack, D.A. Fishman. – N.Y. : Springer, 2009. – 409 p.

7. Bamberger, E. Angiogenesis in epithelian ovarian cancer (review) / E. Bamberger, C. Perrett // Molecular Pathology. – 2002. – № 55. – P. 348–359.

8. Computer-aided image processing of angiogenic histological samples in ovarian cancer / M. Sprindzuk [et al.] // Journal of Clinical Medicine Research. – 2009. – Vol. 1, № 5. – P. 249–261.

9. Folkman, J. What is the evidence that tumors are angiogenesis dependent? / J. Folkman // Journal of the National Cancer Institute. – 1990. – Vol. 82, № 1. – P. 4–6.

10. Hsu, W. Image mining: Trends and developments / W. Hsu, M. Lee, J. Zhang // Journal of Intelligent Information Systems. – 2002. – Vol. 19, № 1. – P. 7–23.

11. Herold, J. Multivariate image mining / J. Herold, C. Loyek, T.W. Nattkemper // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. – 2011. – Vol. 1, № 1. – P. 2–13.

12. Perner, P. Image mining: Issues, framework, a generic tool and its application to medical image diagnosis / P. Perner // Engineering Applications of Artificial Intelligence. – 2002. – Vol. 15, № 2. – P. 205–216.

13. Kovalev, V. Mining lung shape from x-ray images / V. Kovalev, A. Prus, P. Vankevich // Machine Learning and Data Mining in Pattern Recognition (MLDM-2009). – Germany, 2009. – Vol. 5632. – P. 554–568.

14. Kovalev, V. Histological image mining for exploring textural differences in cancerous tissue / V. Kovalev, I. Safonau, A. Prus // Swedish Symposium on Image Analysis (SSBA-2010). – Sweden, 2010. – P. 113–116.

15. Image indexing using color correlograms / J. Huang [et al.] // IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition. – USA, 1997. – P. 762–768.

16. Kovalev, V. Color co-occurrence descriptors for querying-by-example / V. Kovalev, S. Volmer // Int. Conf. on Multimedia Modelling. – Switzerland, 1998. – P. 32–38.

17. Julesz, B. Foundations of Cyclopean Perception / B. Julesz. – Cambridge, Massachusetts : The MIT Press, 2006. – 426 p.

18. Cortical regions involved in visual texture perception: a fMRI study / L.L. Beason-Held [et al.] // Cognitive Brain Research. – 1998. – № 7. – P. 111–118.

19. Petrou, M. Three-dimensional nonlinear invisible boundary detection / M. Petrou, V. Kovalev, J. Reichenbach // IEEE Trans. Image Processing. – 2006. – Vol. 15, № 10. – P. 3020–3032.

20. Kovalev, V. Detection of structural differences between the brains of schizophrenic patients and controls / V. Kovalev, M. Petrou, J. Suckling // Psychiatry Research: Neuroimaging. – 2003. – № 124. – P. 177–189.

21. Heckbert, P. Color image quantization for frame buffer display / P. Heckbert // Proc. of the 9th annual conf. on computer graphics and interactive techniques (SIGGRAPH '82). – USA, 1982. – P. 297–307.

Поступила 10.02.12

Объединенный институт проблем
информатики НАН Беларуси,
Минск, Сурганова, 6
e-mail: vassili.kovalev@gmail.com,
dmitruk@newman.bas-net.by

V.A. Kovalev, A.A. Dmitruk, I.U. Safonau

A METHOD FOR IDENTIFICATION AND VISUALIZATION OF HISTOLOGICAL IMAGE STRUCTURES RELEVANT TO THE CANCER PATIENT CONDITIONS

A method is suggested for identification and visualization of histology image structures relevant to the key characteristics of the state of cancer patients. The method is based on a multistep procedure which includes calculating image descriptors, extracting their principal components, correlating them to known object properties and mapping disclosed regularities all the way back up to the corresponding image structures they found to be linked with. As a result, a number of associations between the patients' conditions and morphological image structures were found.