

ПОИСК ПО СОДЕРЖАНИЮ КАК БАЗОВЫЙ ИНСТРУМЕНТ АНАЛИЗА БОЛЬШИХ МЕДИЦИНСКИХ ДАННЫХ



В.А. Ковалев

Заведующий лабораторией анализа биомедицинских изображений ОИПИ НАН Беларуси, кандидат технических наук,
Республика Беларусь



А.А. Калиновский

Научный сотрудник лаборатории анализа биомедицинских изображений ОИПИ НАН Беларуси,
Республика Беларусь



В.А. Левчук

Младший научный сотрудник лаборатории анализа биомедицинских изображений ОИПИ НАН Беларуси,
Республика Беларусь



А.А. Дмитрук

Младший научный сотрудник лаборатории анализа биомедицинских изображений ОИПИ НАН Беларуси,
Республика Беларусь

Лаборатория анализа биомедицинских изображений, Объединенный институт проблем информатики Национальная академия наук Беларуси, vassili.kovalev@gmail.com

Технология поиска изображений по содержанию (Content-based image retrieval, CBIR) используется в задачах компьютерного анализа и распознавания медицинских изображений начиная с 80-х годов прошлого века. С момента появления, данный подход был достаточно хорошо изучен и опробован на различных типах медицинских изображений. Было разработано большое количество способов описания содержимого как 2D так и 3D снимков, предложены эффективные способы быстрого поиска и индексации больших коллекций изображений. Параллельно с этим за последние годы были накоплены большие объемы цифровых медицинских изображений. Увеличение объема данных позволяет свести многие задачи диагностики на основе медицинских изображений к задаче «поиска похожего случая» в уже имеющейся базе данных. В качестве эффективного инструмента может выступать технология поиска по содержанию.

Технология поиска изображений по содержанию

Методы поиска изображений по содержанию (Content-based image retrieval, CBIR) включают в себя набор алгоритмов для эффективного решения задачи поиска “похожих” изображений используя лишь цифровое описание визуальных данных, без привлечения дополнительной априорной информации, такой, например, как текстовое описание. Общая схема технологии представлена на рисунке 1 и включает в себя два основных этапа:

- подготовка базы данных дескрипторов цифровых изображений. Данный этап включает в себя вычисление некоторого компактного цифрового описания для изображений (далее - дескриптора) и выполнение процедуры индексации для ускорения решения задачи поиска;

- операция поиска похожих изображений в базе по заданному пользователем образцу.

Критерий «похожести» определяется типом данных и требованием к задаче поиска. Так, в случае решения задачи медицинской диагностики, критерий «похожести» должен определять сходство в диагнозе, а при решении таких задач как создания медицинского атласа, критерий «похожести» может определять визуальное сходство отдельных частей изображения.

Существует большое количество разработанных типов дескрипторов для задач поиска по содержанию [1]. В случае медицинских изображений главной особенностью является «неявность» и невозможность формализации описания визуальных признаков. Поэтому в качестве дескрипторов медицинских изображений широко используются признаки, построенные на основе статистических распределений локальных характеристик изображений [2, 3].

Для решения проблемы быстрого поиска в больших архивах медицинских изображений используются такие алгоритмы как: иерархическое разбиение пространства признаков [4], семантическое [5] и локально-чувствительное хеширование [6].

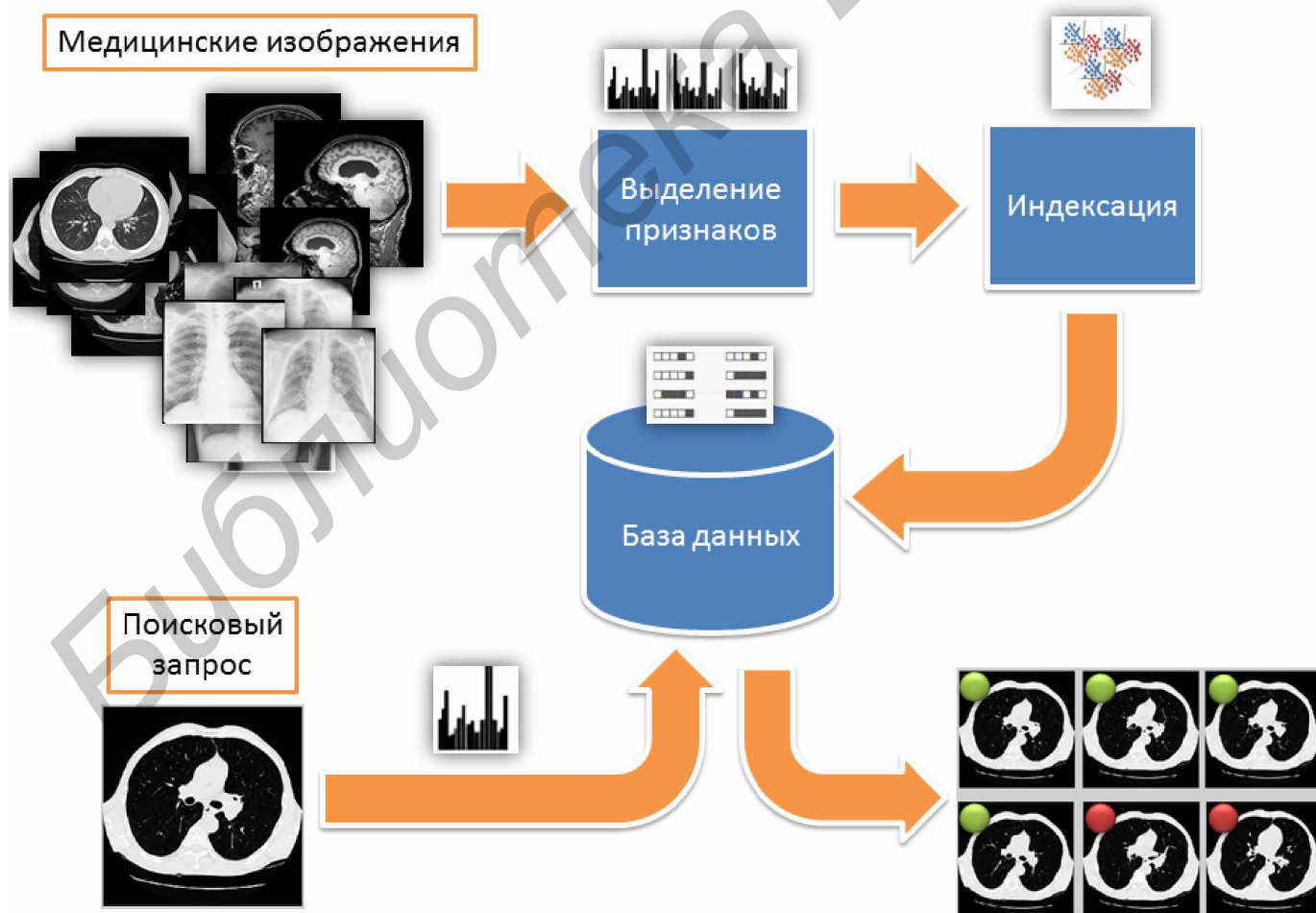


Рис. 2. Общая схема системы поиска медицинских изображений по содержанию

Поиск по содержанию в случае больших данных

В общем случае, за точность поиска по содержанию отвечает выбранный инженером тип дескриптора изображения. При этом увеличить разделяющую силу дескриптора можно за счет увеличения его размера, вычислив какие-то дополнительные характеристики. Однако, увеличивая размер дескриптора, мы тем самым увеличиваем разреженность имеющихся данных в пространстве признаков. Соответственно, тем самым уменьшается вероятность того, что рядом с имеющимся объектом в пространства признаков найдется много сходных.

Решить эту проблему позволяет увеличение количества данных в системе поиска (см. рисунок 2 в качестве пояснения). Следует отметить, что при недостаточном количестве данных пространство признаков оказывается разреженным и вероятность найти похожие экземпляры объектов невелика. При увеличении количества данных в используемой базе вероятность корректного поиска соответственно увеличивается. Таким образом, использование больших объемов данных с эффективным алгоритмом поиска по этим данным позволяет свести многие задачи анализа и обработки медицинских изображений к задаче поиска похожих случаев.

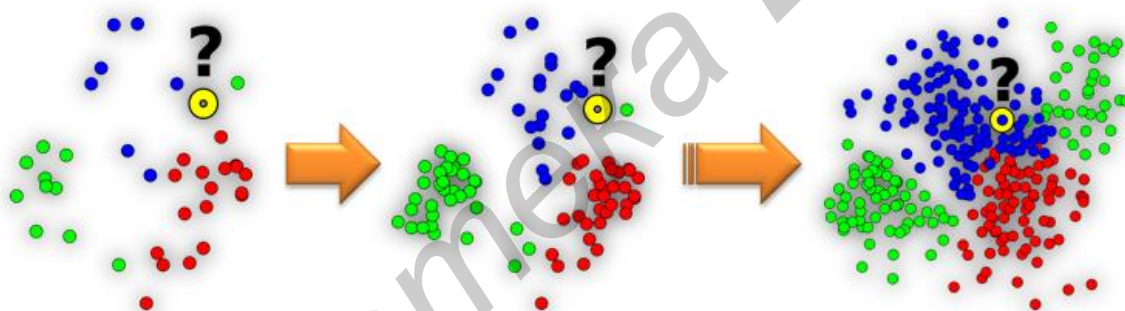


Рис. 3. Изменение результатов поискового запроса по содержанию при увеличении объема базы данных.

Примеры решения задач с использованием концепции поиска по содержанию

Диагностика кожных заболеваний. Классические алгоритмы диагностики злокачественных новообразований по изображению, снятому с подозрительного участка кожи подразумевают сложный анализ как калориметрических, так и морфологических характеристик данного участка (см. так называемый ABCDE-тест [8]). Однако, при наличии базы изображений, представляющей различные случаи, диагностика заболевания может быть сведена к статистическому голосованию по результатам поиска наиболее похожих случаев (см. рисунок 3 и пример web-приложения на сайте [7]).

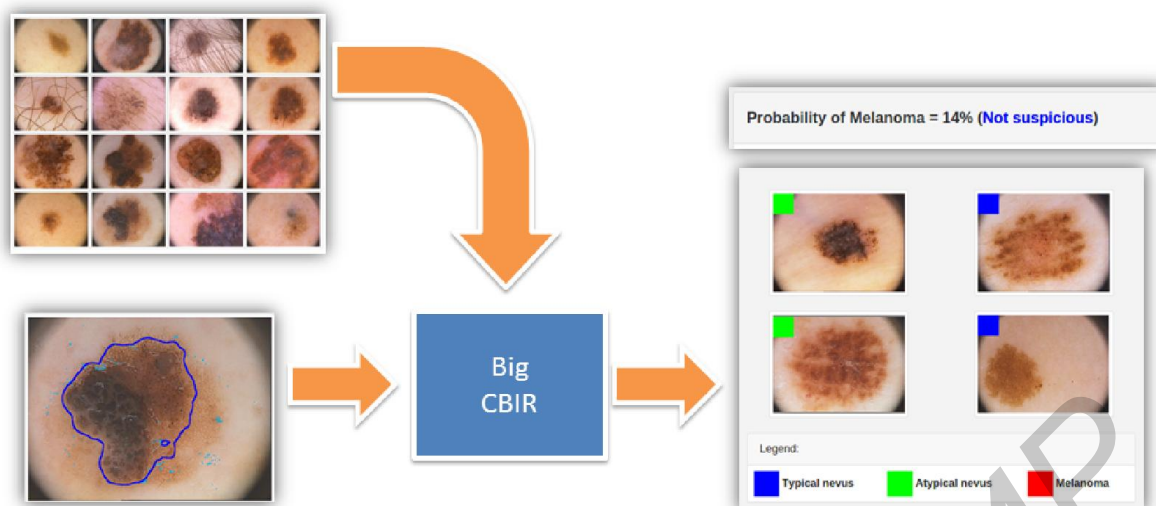


Рис. 4. Использование технологии поиска по содержанию в базе данных цифровых изображений для диагностики кожных заболеваний.

Сегментация области легкого на рентгеновских изображениях грудной клетки [7]. Задача сегментации области легких на рентгеновском изображении грудной клетки является одной из базовых задач в области обработки медицинских изображений. При этом существует большое число алгоритмов для решения данной задачи. Каждый из этих алгоритмов имеет свои преимущества и недостатки [9]. Следует отметить, что эффективность работы большинства методов зависит от первоначальной инициализации потенциальной области расположения легких. Тем не менее, задачу первоначальной инициализации можно эффективно решить простым поиском похожих отсегментированных экземпляров изображений, содержащихся в собранной базе данных (см. рисунок 4 и пример web-приложения на сайте [7], иллюстрируемый рисунком 5).

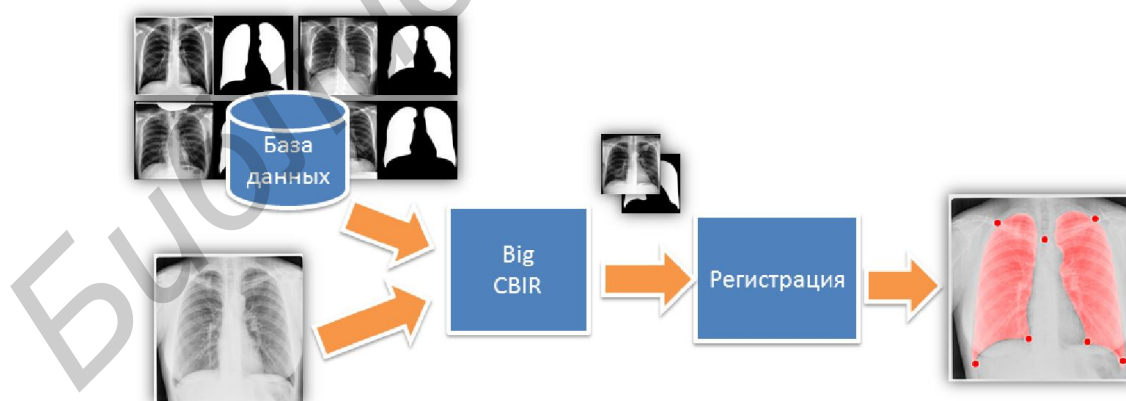


Рис. 5. Использование поиска по содержанию в большой базе данных на предварительном этапе сегментации изображений

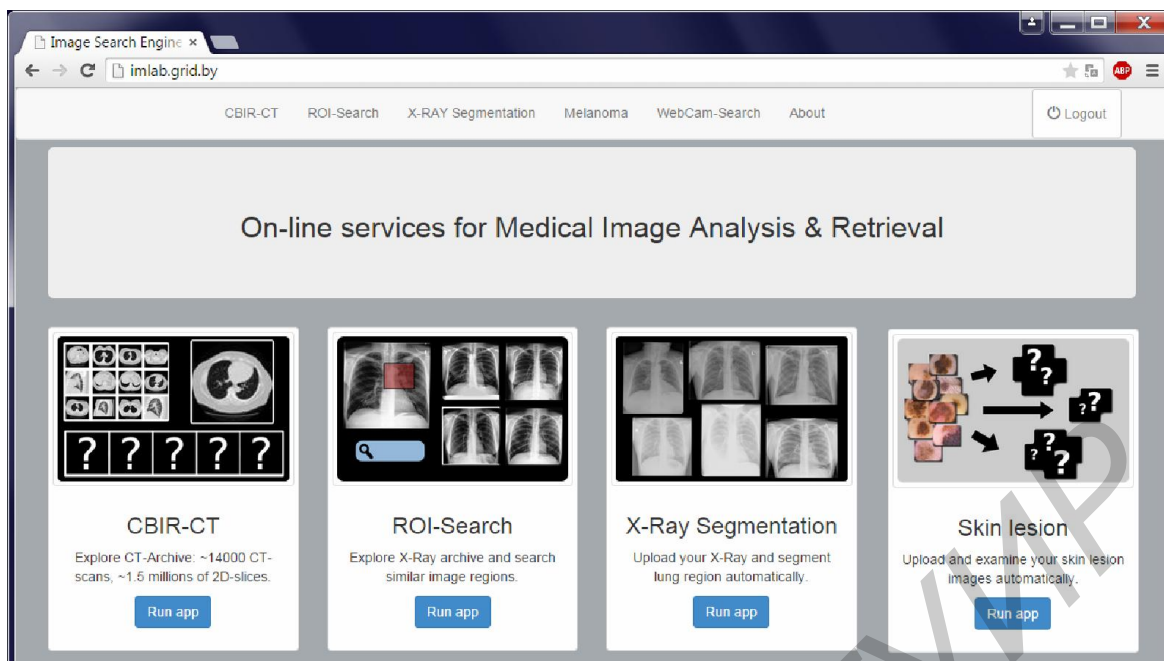


Рис. 6. Экспериментальный сайт веб-сервисов по поиску и анализу медицинских изображений.

Отобрав потенциальных кандидатов, можно далее воспользоваться каким-либо алгоритмом уточнения границ легкого (например, используя метод нежесткой регистрации [10]).

Литература

1. Lew M. S. et al. Content-based multimedia information retrieval: State of the art and challenges //ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). – 2006. – Т. 2. – №. 1. – С. 1-19.
2. Kovalev V., Petrou M. Multidimensional co-occurrence matrices for object recognition and matching //Graphical Models and Image Processing. – 1996. – Т. 58. – №. 3. – С. 187-197.
3. Wang J. et al. Bag-of-features based medical image retrieval via multiple assignment and visual words weighting //IEEE transactions on medical imaging. – 2011. – Т. 30. – №. 11. – С. 1996-2011.
4. Nister D., Stewenius H. Scalable recognition with a vocabulary tree //Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. – IEEE, 2006. – Т. 2. – С. 2161-2168.
5. Salakhutdinov R., Hinton G. Semantic hashing //RBM. – 2007. – Т. 500. – №. 3. – С. 500.
6. Andoni A., Indyk P. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions //Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on. – IEEE, 2006. – С. 459-468.
7. <http://imlab.grid.by/> last visited 16 May 2015.
8. Friedman R. J., Rigel D. S., Kopf A. W. Early detection of malignant melanoma: The role of physician examination and self - examination of the skin //CA: a cancer journal for clinicians. – 1985. – Т. 35. – №. 3. – С. 130-151.
9. Wang C., Guo S., Wu X. Segmentation of Lung region for chest X-ray images based on medical registration and ASM //Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on. – IEEE, 2009. – С. 1-4.
10. Klein S. et al. Elastix: a toolbox for intensity-based medical image registration //Medical Imaging, IEEE Transactions on. – 2010. – Т. 29. – №. 1. – С. 196-205.