

Выбор опорных точек при построении vp-деревьев для поиска дубликатов на цифровых изображениях

Е.В. Мясников^{1,2}

¹Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34а, Самара, Россия, 443086

²Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001

Аннотация

В рамках решения задачи поиска фрагментов-дубликатов на цифровых изображениях рассматривается несколько способов выбора опорных точек при построении vp-деревьев, используемых в работе для ускорения поиска.

Ключевые слова

Duplicate, Copy-move, Forgery, Vp-tree, Binary space partitioning tree

1. Введение

Стандартный подход к поиску фрагментов-дубликатов на цифровых изображениях состоит в расчете признаков в скользящем окне с последующим поиском ближайших фрагментов в признаковом пространстве. Учитывая возможные размеры анализируемых изображений и размерность пространства признаков, поиск ближайших занимает значительное время.

Для повышения эффективности поиска могут использоваться древовидные структуры разбиения многомерного пространства признаков, такие как kd-деревья и vp-деревья. Последние были рассмотрены в работе [1] и показали хорошие результаты при решении рассматриваемой задачи. При этом vp-деревья допускают использование различных схем выбора опорных точек, что и является предметом исследования в настоящей работе.

2. Методы и результаты

Vp-дерево [2] представляет собой бинарное дерево, разбивающее многомерное пространство признаков с использованием гиперсфер. Разбиение производится в рекурсивном порядке. На каждом этапе из множества точек S некоторым образом выбирается одна опорная точка $v \in S$, и все множество точек разбивается на два одинаковых по мощности множества S_1 и S_2 :

$$S_1 = \{p \in S \setminus \{v\} \mid d(p, v) < d_m\}, \quad (1)$$

$$S_2 = \{p \in S \setminus \{v\} \mid d(p, v) > d_m\},$$

где d_m – медианное значение расстояний $\{d(p, v), p \in S \setminus \{v\}\}$.

Можно предположить, что способ выбора опорных точек может существенным образом влиять как на время построения, так и на эффективность использования vp-деревьев. В рамках настоящей работы исследуются следующие способы выбора опорных точек:

- Стандартный подход. Формируется два случайных подмножества точек. Первое из подмножеств содержит потенциальных кандидатов в опорные точки. Второе – тестовые точки, по которым оцениваются кандидаты. В качестве опорной точки выбирается та точка-кандидат, для которой максимальна дисперсия расстояний до точек из тестового множества. (далее подход обозначен как $RS-n$, где n – мощность случайных подмножеств).
- В качестве опорной выбирается точка, наиболее удаленная от центра (среднего), рассчитанного по всему множеству точек (обозначено FFC).
- Выбирается любая (в текущей реализации всегда первая) точка (обозначено FP).
- Выбирается точка, вектор координат которой имеет наибольшую длину (обозначено LV).

- Выбирается точка с наибольшей по модулю проекцией на первую главную компоненту (обозначено $PC-n$, где n – количество точек, взятых для оценки главной компоненты).

Для оценки указанных способов выбора опорных точек использовался набор данных, описанный в [1]. Исследования проводились с использованием ПК на базе Intel Core i3-6100U CPU @ 2.3 ГГц. Результаты исследования представлены на Рисунке 1.

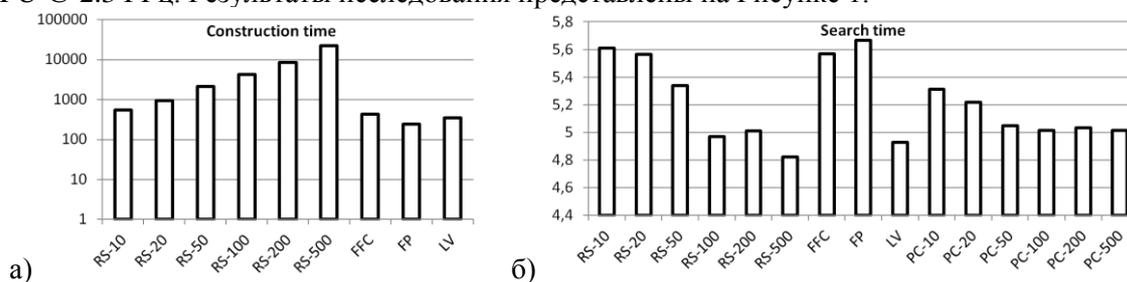


Рисунок 1: Среднее время построения (а) и время поиска (б) в миллисекундах для vr- деревьев, построенных с использованием различных подходов

Как можно было ожидать, наименьшее время при построении (Рисунок 1.а, шкала времени логарифмическая) показывает способ FP , не требующий для выбора опорной точки даже однократного прохода по списку. Этому же способу соответствует наибольшее время поиска (Рисунок 1.б). Способы LV и FFC демонстрируют несколько большее время построения и требуют одно- и двукратного прохода по множеству для выбора опорной точки. При этом LV демонстрирует еще и отличные результаты при поиске, уступая лишь стандартному подходу при использовании сравнительно больших подмножеств ($RS-500$ на рисунке). Способ FFC показывает существенно большее время поиска. Стандартный подход (RS) и способ на основе оценки первой главной компоненты (PC) позволяют в некоторой степени выбрать компромисс между временем поиска и временем построения деревьев. Тем не менее, те значения, при которых достигается сравнимое с LV или лучшее время поиска характеризуются заметно большим, чем у LV , временем построения (ввиду ограниченного объема время построения для PC не приводится и превышает таковое для способа RS при сравнимых параметрах).

3. Заключение

Рассмотрен ряд способов выбора опорных точек при построении vr-дерева, используемого в работе для ускорения поиска фрагментов-дубликатов на цифровых изображениях. Показано, что RS и LV подходы обеспечивали минимальное время поиска, в то время как LV подход показал и сравнительно малое время построения структуры данных. Планируется расширение исследований с использованием альтернативных пространств признаков и наборов данных.

4. Благодарности

Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта № 20-37-70053 в частях «2. Методы и результаты» - «3. Заключение» и Министерства науки и высшего образования РФ в рамках госзадания ФНИЦ «Кристаллография и фотоника» РАН в части «1. Введение».

5. Литература

- [1] Kuznetsov, A. Copy-move detection algorithm efficiency increase using binary space partitioning trees / A. Kuznetsov, E. Myasnikov // CEUR Workshop Proceedings. – 2016. – Vol. 1638. – P. 373-378.
- [2] Yianilos, P. Data structures and algorithms for nearest neighbor search in general metric spaces / P. Yianilos // Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms. – 1993. – P. 311-321.