

УДК 514.86

DOI: 10.20998/2411-0558.2018.24.02

А. А. ДАШКЕВИЧ, канд. техн. наук, доц., докторант, НТУ "ХПИ",
О. В. ШОМАН, д-р. техн. наук, проф., зав. каф., НТУ "ХПИ"

АНАЛИЗ ПРОСТРАНСТВЕННОГО РАСПРЕДЕЛЕНИЯ ТОЧЕЧНЫХ МНОЖЕСТВ НА ОСНОВЕ АЛГОРИТМА ПРОСТРАНСТВЕННОГО ХЕШИРОВАНИЯ

В работе предложен алгоритм для определения меры пространственного распределения точечных множеств на основе алгоритма пространственного хеширования. В работе вводится расширение меры пространственной плотности точечного множества на трехмерное пространство. Разработанный подход позволяет выявлять корреляцию двух точечных множеств как скалярного произведения распределения плотностей индексов ближайшего соседства, что дает возможность решения задачи сегментации точечных множеств. Ил.: 2. Библиогр.: 15 назв.

Ключевые слова: точечное множество; пространственное хеширование; мера пространственной плотности; индекс ближайшего соседства; сегментация точечных множеств.

Постановка проблемы. В задачах компьютерного зрения данные с сенсоров камеры часто представлены в виде точечных множеств: матрицы яркости пикселей при фотосъемке, облака точек в задачах стереозрения. При этом возникает необходимость сегментации таких точечных множеств, т. е. их разбиения на независимые компоненты. Задача сегментации предполагает решение двух связанных подзадач:

1. Нахождение точек, которые находятся в некоторой окрестности от заданной;

2. Нахождение геометрических характеристик точечного множества в локальных окрестностях.

На данный момент точное решение для задачи поиска ближайших соседей дают переборные алгоритмы поиска, которые обладают квадратичной вычислительной сложностью. Выделение геометрических признаков в локальной окрестности основано на анализе пространственного распределения точек и статистических показателях ближайшего соседства. Таким образом, качество определения геометрических признаков тесно связано с эффективностью определения ближайших соседей точек. В связи с этим существует необходимость разработки алгоритмов определения геометрических характеристик на основе пространственного анализа с достаточной точностью и вычислительной сложностью в пространствах произвольной размерности.

Анализ последних исследований. Для поиска ближайших соседей используются специальные структуры данных, например kd -деревья [1, 2], BSP -деревья [3, 4], R -деревья [5] и др., а также алгоритмы по обработке таких структур [6]. Однако, эффективность подобных структур данных и алгоритмов падает при росте числа измерений N [7]. В настоящее время развиваются алгоритмы приближенного поиска ближайших соседей в многомерных пространствах, среди которых можно выделить методы, основанные на использовании хеширующих функций, которые однозначно указывают на область пространства, которой принадлежит точка.

В работе [8] представлен метод пространственного хеширования, в котором предлагается заменить координаты x , y и z точек множества на одну координату – хеш, который предлагается вычислять по следующему соотношению:

$$h = x/c \cdot 2^k + y/c \cdot 2^m + z/c \cdot 2^n,$$

где c – размер сетки; k , m и n – некоторые константы, такие что, $k > m > n$ (или $k < m < n$).

В работе [9] предложен следующий подход для вычисления хеша:

$$h = ((x \cdot p_1) \text{ XOR } (y \cdot p_2) \text{ XOR } (z \cdot p_3)) \text{ MOD } n,$$

где $p_1 = 73856093$, $p_2 = 19349663$, $p_3 = 83492791$ – большие простые числа, n – общее число точек.

В работе [10] предложен алгоритм приближенного поиска ближайших соседей на двумерной сетке для применения в решении задач молекулярной динамики.

Одним из преимуществ данного алгоритма является отсутствие коллизий хешей. Однако, данный алгоритм может быть использован для пространств с $N = 2$, что приводит к необходимости его расширения на пространства большей размерности.

В работах [11, 13] изучаются алгоритмы и структуры данных для поиска ближайших соседей в многомерных пространствах. В [13 – 15] рассматриваются различные подходы по выявлению геометрической структуры на плоскости и в пространстве. Работа [15] посвящена пространственным статистикам для задач географии.

Цель работы. Проведение пространственного анализа N -мерных точечных множеств на основе алгоритма упрощенного пространственного хеширования и меры плотности пространственного распределения множества.

Основная часть. В работе предлагается следующее расширение алгоритма, предложенного в работе [10] для пространственного хеширования в пространствах с $N > 2$:

1) пространство разбивается на многомерную сетку с гиперкубическими ячейками с длиной ребра гиперкуба c ;

2) для каждой точки с координатами $(x_1, \dots, x_i, \dots, x_N)$, заданными на интервале $[0, G_i]$ ($G_i \geq c$ – максимальное значение для координатной оси) вычисляются индексы клетки, в которой находится данная точка:

$$i_i = \frac{x_i}{c}, \dots, i_N = \frac{x_N}{c},$$

где N – размерность пространства;

3) по заданным индексам вычисляются значения хеша:

$$h_k = \frac{i_i \cdot 10^{(N-t)d}}{N},$$

где d – максимальное число разрядов хеша, которое приходится на одну пространственную координату x_i ,

$$t = \max\left(\frac{G_i}{c}\right).$$

Был разработан алгоритм разбиения полученного хеша на отдельные индексы по каждой из координатных осей:

$t = N$ <p>до тех пор, пока $t > 0$, повторять:</p> $i_i = h \bmod 10^d$ $h = h / 10^d$ $t = t - 1$

Данный алгоритм разбивает пространство на многомерную регулярную сетку и назначает каждой ячейке сетки свой уникальный номер-хеш, который явным образом зависит от координат ограничивающего гиперкуба ячейки.

Для изучения пространственного распределения в ячейках предлагается расширить подход из [15] на трехмерное пространство:

1) расчет плотности точек в i -й ячейке:

$$p_i = \frac{n_i}{V},$$

где n_i – количество точек в ячейке; $V = c^N$ – объем ячейки (одинаковый для всех ячеек);

2) расчет среднего расстояния между парами ближайших точек в ячейках:

$$\tilde{d}_i = \frac{\sum_{j=1}^{N_P} d_j}{N_P},$$

где N_P – число пар точек в ячейке; d_j – расстояние между парами;

3) ожидаемое среднее (индекс случайного распределения):

$$d_E = \frac{1}{\sqrt[3]{p_i}},$$

4) индекс ближайшего соседства:

$$R_i = \frac{\tilde{d}_i}{d_E}.$$

Значение индекса ближайшего соседства позволяет делать выводы о распределении точек множества в пределах ячейки: при $R < 1$ точки в ячейке рассеяны, при $R = 1$ – распределены случайно, при $R > 1$ – сгруппированы в кластеры.

В рамках исследования поиск всех пар ближайших точек заменен на случайно выбранные пары точек в пределах ячейки, что может быть эффективно при вычислениях в пространствах большой размерности.

Для проверки приведенных положений был проведен расчет хешей и индексов ближайшего соседства для двух точечных множеств, в качестве которых были использованы данные карт глубины двух различных трехмерных сцен (рис. 1). На картах глубины пространственные координаты x и y соответствуют ширине и высоте изображения, координата z кодируется в виде значения уровня яркости пикселей изображения от 0 (черный цвет) до 255 (белый цвет). На рис. 2 приведены соответствующие хеш-разбиения исходных точечных множеств для $N = 3$ с учетом индекса ближайшего соседства, который показан как диаметр точки. При этом пространственной оси x исходной карты глубин соответствует ось пространственных индексов i_1 , оси y – ось i_2 , оси z – ось i_3 хеш-разбиения. Из анализа разбиений и индексов ближайшего соседства можно увидеть, что точки, соответствующие каждому конкретному объекту исходной сцены имеют близкие значения индекса ближайшего соседства, что позволяет проводить разбиение

точечного множества на отдельные подобласти путем пороговой обработки по критическим значениям таких индексов.

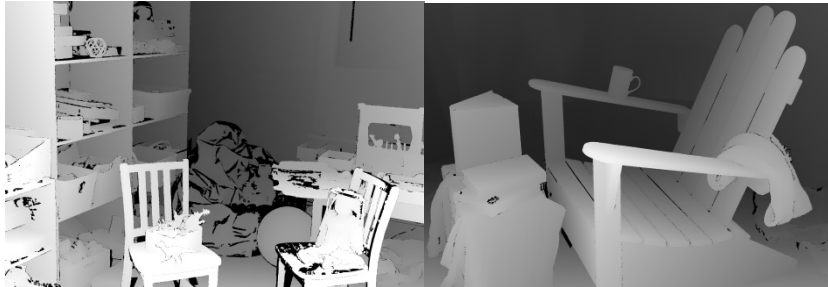


Рис. 1. Карты глубины

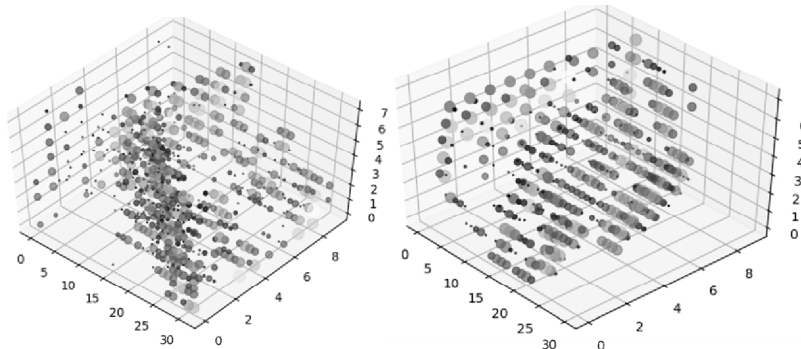


Рис. 2. Хеши множеств с учетом индекса ближайшего соседства

Предложенный подход позволяет оценить глобальную геометрическую структуру точечного множества, однако не позволяет напрямую сравнивать точечные множества между собой. Для выявления сходства точечных множеств предлагается способ вычисления коэффициента корреляции множеств как скалярного произведения нормированных плотностей распределения индексов ближайшего соседства. Для указанных множеств были вычислены гистограммы распределения плотностей и вычислено значение коэффициента корреляции, которое составило $C \approx 0.36$, что позволяет делать выводы о мере различия данных точечных множеств.

Выводы и перспективы дальнейших исследований. Научную новизну предложенной работы составляют:

1. Алгоритм пространственного хеширования для быстрого вычисления приблизительных ближайших соседей точек в N -мерном пространстве.

2. Способ вычисления пространственных характеристик точечных множеств на основе индекса ближайшего соседства.

3. Способ вычисления меры сходства двух множеств как скалярного произведения нормированных плотностей распределения индексов ближайшего соседства.

Практическая ценность подхода заключается в возможности его применения при решении таких задач компьютерного зрения и геометрического моделирования, как сопоставление и объединение карт глубины, классификация и сегментация точечных множеств.

Дальнейшие исследования будут направлены на выделение локальных характеристик точечных множеств и их топологическую сегментацию.

Список литературы:

1. Bentley J.L. Multidimensional Divide and Conquer / J.L. Bentley // Communications of the ACM, 1980. – Vol. 23. – №. 4. – P. 214-229.
2. Friedman J.H. An algorithm for finding best matches in logarithmic expected time / J.H. Friedman, J.L. Bentley, and R.A. Finkel // ACM Transactions on Mathematical Software, 1977. – Vol. 3. – №. 3. – P. 209-226.
3. de Berg M. Computational Geometry: Algorithms and Applications / M. de Berg // Springer Science & Business Media, 2008. – 259 p.
4. Castelli V. Image Databases: Search and Retrieval of Digital Imagery / V. Castelli, D. Lawrence // John Wiley & Sons, 2004. – 422 p.
5. Guttman A. R-Trees: A Dynamic Index Structure for Spatial Searching / A. Guttman // Proceedings of ACM SIGMOD International Conference on Management of Data, 1984. – P. 47-57.
6. Liu T. An Investigation Of Practical Approximate Nearest Neighbor Algorithms / T. Liu, A. Moore, A. Gray, K. Yang // Neural Information Processing Systems, 2004. – P. 825-832.
7. Beyer K. When Is "Nearest Neighbor" Meaningful? / K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft // In International Conference on Database Theory: Springer, 1999. – P. 217-235.
8. Nguyen H. (ed.). GPU Gems 3 / H. Nguyen // Addison Wesley, 2007. – 1008 p.
9. Ize T. Asynchronous BVH Construction for Ray Tracing Dynamic Scenes on Parallel Multi-Core Architectures / T. Ize, I. Wald, S.G. Parker // Proceedings of the 2007 Eurographics Symposium on Parallel Graphics and Visualization, 2007. – P. 101-108.
10. Ніцин О.Ю. Спрощений алгоритм просторового хешування для задач молекулярної динаміки / О.Ю. Ніцин, А.О. Дашкевич, О.В. Охотська, О.Є. Мацулевич // Науковий вісник таврійського державного агротехнологічного університету. – Мелітополь: ТДАТУ, 2016. – Вип. 6. – Т. 1. – С. 287-291.
11. Muja M. Scalable Nearest Neighbor Algorithms for High Dimensional Data / M. Muja, D.G. Lowe // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014. – Vol. 36. – P. 2227-2240.
12. Megala, R. Hash tree algorithm for centralized cognitive radio networks / R. Megala, S. Sathya // International Journal of Computer Science and Mobile Computing, 2014. – Vol. 3. – № 11. – P. 305-315.
13. Bewley A. Advantages of exploiting projection structure for segmenting dense 3D point clouds / A. Bewley, B. Upcroft // Proceedings of the 2013 Australasian Conference on Robotics and Automation, Australian Robotics & Automation Association, University of New South Wales, Sydney, NSW, 2013. – P. 1-8.

14. Li L. Point Pattern Analysis for Clusters Influenced by Linear Features: An Application for Mosquito Larval Sites / L. Li, L. Bian, P. Rogerson, G. Yan // *Transactions in GIS*, 2015. – Vol. 19. – 835-847.

15. McGrew J. An introduction to statistical problem solving in geography, 2nd ed. / J. McGrew, C. Monroe // Boston : McGraw-Hill, 2000. – 254 p.

References:

1. Bentley, J.L., (1980). "Multidimensional Divide and Conquer". *Communications of the ACM*, 1980, Vol. 23, No. 4, pp. 214-229.

2. Friedman, J.H., Bentley, J.L. and Finkel R.A., (1977). "An algorithm for finding best matches in logarithmic expected time". *ACM Transactions on Mathematical Software*, 1977, Vol. 3, No. 3, pp. 209-226.

3. de Berg, M., (2008). *Computational Geometry: Algorithms and Applications*. Springer Science & Business Media, 2008, 259 p.

4. Castelli, V., and Lawrence, D., (2004). *Image Databases: Search and Retrieval of Digital Imagery*. John Wiley & Sons, 422 p.

5. Guttman, A. (1984). "R-Trees: A Dynamic Index Structure for Spatial Searching". *Proceedings of ACM SIGMOD International Conference on Management of Data*, 1984, pp. 47-57.

6. Liu, T., Moore, A., Gray, A., and Yang, K., (2004). "An Investigation Of Practical Approximate Nearest Neighbor Algorithms". *Neural Information Processing Systems*, Vol. 2, No. 3, pp. 825-832.

7. Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U., (1999). "When Is "NearestNeighbor" Meaningful?" *In International Conference on Database Theory: Springer*, pp. 217-235.

8. Nguyen, H., (2007). *GPU Gems 3*. Addison Wesley, 2007, 1008 P.

9. Ize, T., Wald, I. and Parker S.G. (2007). "Asynchronous BVH Construction for Ray Tracing Dynamic Scenes on Parallel Multi-Core Architectures". *Proceedings of the 2007 Eurographics Symposium on Parallel Graphics and Visualization*, pp. 101-108.

10. Nitsyn, A., Dashkevich, A., Ohotskaya, E., and Matsulevich A., (2016). "Simplified spatial hashing algorithm for molecular dynamics problems". *Scientific bulletin of the Tavria agrotechnological state university*, No. 6, Vol. 1, pp. 287-291.

11. Muja, M., and Lowe, D.G., (2014). "Scalable Nearest Neighbor Algorithms for High Dimensional Data". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 36, pp. 2227-2240.

12. Megala, R., and Sathya, S. (2014). "Hash tree algorithm for centralized cognitive radio networks". *International Journal of Computer Science and Mobile Computing*, Vol. 3, No. 11, November, pp. 305-315.

13. Bewley, A., and Upcroft, B., (2013). "Advantages of exploiting projection structure for segmenting dense 3D point clouds". *Proceedings of the 2013 Australasian Conference on Robotics and Automation*, University of New South Wales, Sydney, NSW, pp. 1-8.

14. Li, L., Bian, L., Rogerson, P., and Yan, G., (2015). "Point Pattern Analysis for Clusters Influenced by Linear Features: An Application for Mosquito Larval Sites". *Transactions in GIS*, Vol. 19, pp. 835-847.

15. McGrew, J., and Monroe, C., (2000). *An introduction to statistical problem solving in geography*. 2nd ed. McGraw-Hill, Boston, 254 p.

*Статью представил д-р техн. наук, проф. Национального
технического университета "Харьковский политехнический институт"
А.Ю. Ницын*

Поступила (received) 15.04.2018

Повторно 05.05.2018

Dashkevich Andrey, Cand. Tech, Sci.
National Technical University "KhPI",
Str. Кирпичова, 2, Kharkiv, Ukraine, 61002
Tel.: +38 (095) 388-04-56, e-mail: dashkevich.a@gmail.com
ORCID ID:0000-0002-9963-0998

Shoman Olga, Doct. Tech. Sci.
National Technical University "KhPI",
Str. Кирпичова, 2, Kharkiv, Ukraine, 61002
Tel.: +38 (050) 806-94-74, e-mail: olgasho@ukr.net
ORCID ID:0000-0002-3660-0441

УДК 514.86

Аналіз просторового розподілу точкових множин на основі алгоритму просторового хешування / Дашкевич А.О., Шоман О.В. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2018. – № 24 (1300). – С. 16 – 24.

В роботі запропоновано алгоритм для визначення міри просторового розподілу точкових множин на основі алгоритму просторового хешування. В роботі вводиться розширення міри просторової густини точкової множини на тривимірний простір. Розроблений підхід дозволяє визначати кореляцію двох точкових множин як скалярного добутку розподілу густини індексів найближчого сусідства, що дозволяє розв'язувати задачі сегментації точкових множин. Ил.: 2. Бібліогр.: 15 назв.

Ключові слова: точкова множина; просторове хешування; міра просторової густини; індекс найближчого сусідства; сегментація точкових множини.

УДК 514.86

Анализ пространственного распределения точечных множеств на основе алгоритма пространственного хеширования / Дашкевич А.А., Шоман О.В. // Вестник НТУ "ХПИ". Серія: Інформатика и моделирование. – Харьков: НТУ "ХПИ". – 2018. – № 24 (1300). – С. 16 – 24.

В работе предложен алгоритм для определения меры пространственного распределения точечных множеств на основе алгоритма пространственного хеширования. В работе вводится расширение меры пространственной плотности точечного множества на трехмерное пространство. Разработанный подход позволяет выявлять корреляцию двух точечных множеств как скалярного произведения распределения плотностей индексов ближайшего соседства, что дает возможность решать задачи сегментации точечных множеств. Ил.: 2. Библиогр.: 15 назв.

Ключевые слова: точечное множество; пространственное хеширование; мера пространственной плотности; индекс ближайшего соседства; сегментация точечного множества.

UDC 514.86

Analysis of spatial distribution of point sets based on spatial hashing algorithm / Dashkevich A.A., Shoman O.V. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2018. – № 24 (1300). – P. 16 – 24.

An algorithm for determining the measure of the spatial distribution of point sets based on the spatial hashing algorithm is proposed. In this paper we extend the measure of the spatial density of a point set to a three-dimensional space. The developed approach allows us to identify the correlation of two point sets as a scalar product of the density distribution of nearest neighborhood index, which makes it possible to solve the problems of segmentation of point sets. Figs.: 2. Refs.: 15 titles.

Keywords: point set; spatial hashing; measure of the spatial density; nearest neighborhood index; point set segmentation.