Оценка возможности восстановления 3D-сцены по последовательности изображений

Е.А. Дмитриев[1], В.В. Мясников[1,2]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086
[2]Институт систем обработки изображений РАН - филиал ФНИЦ «Кристаллография и фотоника» РАН, Молодогвардейская 151, Самара, Россия, 443001

Аннотация. В данной работе предложен метод получения попиксельной оценки "надежности" восстановления 3D-сцены по последовательности изображений. Подход основан на оценке количества сопряжённых пар с использованием свёрточных нейронных сетей. В качестве критериев эффективности работы алгоритма выступают точность получаемой оценки, а также скорость работы алгоритма. Все эксперименты проводились на наборе данных, полученных с помощью программы Unity. Результаты исследований показали эффективность используемого метода в задаче восстановления трёхмерной сцены.

1.Введение

3D-реконструкция сцен является классической задачей компьютерного зрения. Алгоритмы для восстановления трёхмерной сцены используются в таких областях как робототехника, архитектура, дизайн, системы дистанционного зондирование земли, системы автоматического управления транспортными средствами.

Существует несколько алгоритмов для решения рассматриваемой задачи [1,2]. Одним из таких алгоритмов является метод бинокулярного стереовидения [1]. Данный метод основан на нахождении диспаратности между сопряжёнными парами точек двух изображений. Основная проблема такого подхода заключается в нахождении таких пар. Один из возможных методов решения состоит в поиске ключевых точек на изображениях, получении описаний или дескрипторов точек и их сопоставлении на изображениях по значению метрики между дескрипторами [3]. Существуют и более современные методы, основанные на использовании глубоких свёрточных нейронных сетей [4].

Для реконструкции 3D-сцены с использованием последовательности изображений требуется много времени [2]. К сожалению, на данном этапе развития технологий и алгоритмов 3D-реконструкции пока невозможно получать изображения с 3D сценой в режиме реального времени.

В данной статье описывается алгоритм, позволяющий оценивать в режиме реального времени возможность 3D–восстановления сцен по имеющимся кадрам. Суть алгоритма заключается в оценке количества сопряжённых пар по последовательности изображений с использованием глубокой свёрточной нейронной сети. В статье представлена архитектура нейронной сети со сравнительно небольшим количеством параметров. Такую архитектуру можно использовать на

мобильных графических процессах в режиме реального времени. Все эксперименты были проведены на изображениях, полученных с помощью программы Unity.
Работа выстроена в следующем порядке. Во втором разделе приводятся основные понятия. В следующем разделе даётся описание используемого алгоритма. В четвёртом разделе представлены результаты экспериментов. В заключительном разделе подведены итоги экспериментов и сказано про направление дальнейших исследований.

2. Основные понятия

Пусть $I_k^s(n_1,n_2)$ – изображение, полученное с камеры k и сцены s, где $(n_1,n_2)\in\mathbf{D}$, $\mathbf{D}=\{(n_1,n_2): n_1=\overline{0,N_1-1}, n_2=\overline{0,N_2-1}\}$, $k=\overline{0,K-1}$, $s=\overline{0,S-1}$, N_1,N_2 – размеры получаемого с камеры изображения, K – количество камер и S – количество сцен. Пусть $R_k^s(n_1,n_2)$ – дискретная функция, значение которой есть координаты точки объекта в пространстве. Каждому значению $R_k^s(n_1,n_2)$ соответствует проекция точки на плоскости изображения $I_k^s(n_1,n_2)$. Обозначим за l индекс, соответствующий опорному кадру или функции $R_l^s(n_1,n_2)$, значения которой проецируются на опорный кадр. Для формирования элементов обучающей и тестовой выборок определим следующую функцию:

$$P_j^s\left(n_1,n_2\right)=\begin{cases}0, R_j^s\left(n_1,n_2\right)\neq R_l^s\left(n_1,n_2\right)\\ 1, R_j^s\left(n_1,n_2\right)=R_l^s\left(n_1,n_2\right)\end{cases} \quad (1)$$

Пусть $X^G=\left(x_i,y_i\right)_{i=0}^{G-1}$ – обучающая выборка, где x_i – элемент выборки, подаваемый на вход алгоритма, y_i – требуемое значение или ответ алгоритма при предъявлении x_i, а G – размер выборки. В качестве элемента x_i выступает набор из $m<K$ различных изображений одной сцены вместе с опорным кадром. Количество кадров в наборе меньше, чем количество камер с одной сцены для того, чтобы формировать несколько элементов x_i. Таким образом, количество обучающих элементов, получаемых с изображений одной сцены, равняется C_{K-1}^{m-1}.
Для получения ответа y_i формируем следующую вспомогательную функцию:

$$A_i^s(n_1,n_2)=\sum_{\substack{j=0\\ j\neq l}}^{m-1}P_j^s\left(n_1,n_2\right), \quad (2)$$

где j соответствует номеру камеры из набора без учёта камеры, которая даёт опорный кадр. Значение функции $A_i^s\left(n_1,n_2\right)$ показывает количество кадров (без учета опорного кадра) в наборе с m изображениями, для которых реальная точка объекта в пространстве, проецирующая на пиксель $\left(n_1,n_2\right)$ опорного кадра, также проецируется на пиксель других кадров. Ответ y_i равен функции $A_i^s(n_1,n_2)$, где каждое значение кодируется в унитарном коде с m разрядами. В итоге y_i представляет собой изображение с m каналами. Каждый канал соответствует определенному классу. Значение отсчета в канале равняется 0 или 1 в зависимости от количества сопряжённых пар точек в наборе для данного отсчета опорного кадра. Таким образом, описанный ниже алгоритм используется для решения задачи попиксельной классификации.

3. Описание алгоритма

В качестве моделей нейронных сетей рассматривались полносвёрточные нейронные сети, позволяющие на выходе получить карты признаков, размер которых совпадает с размером исходных изображений [5]. Такие сети используются в задаче сегментации. Примерами таких моделей выступают сети U-net [6], SegNet [7]. Несмотря на результаты, которые позволяют получить перечисленные выше нейронные сети, они обладают большим недостатком –

большое количество параметров. Следовательно, на использование таких сетей тратится достаточно много времени, не говоря уже о времени обучения.
Неплохим вариантом является модель LinkNet [8]. Данная модель использует преимущества сети U-net, но обладает меньшим числом параметров и позволяет достигать хорошую точность. Отличительной чертой сети является использование нескольких блоков кодеров и декодеров. Оригинальная модель сети состоит из 4 блоков каждого типа. Количество параметров такой сети равняется 11,5 миллионов.
В данной работе вместо 4 блоков используется 3 блока, что позволяет сократить количество параметров до 3 миллионов. Во втором слое вместо прореживания с ядром 3×3 и шагом 2, используется ядро 2×2 с шагом 1. Отличием предлагаемой архитектуры от оригинальной является количество используемых каналов. В оригинальной архитектуре предполагалось использование 3 каналов цветного изображения. Модель сети представлена на рисунке 1. На рисунка 2 и 3 представлена архитектура кодера и декодера соответственно.

Рисунок 1. Модель используемой нейронной сети.

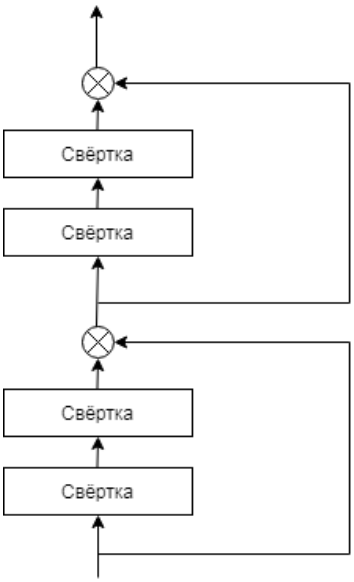

Рисунок 2. Архитектура кодера.

Рисунок 3. Архитектура декодера.

В качестве функции потери выступала перекрёстная энтропия. Согласно [9], в задаче классификации использование перекрёстной энтропии в качестве функции потери позволяет достичь лучшего локального минимума с точки зрения точности классификации при случайной инициализации параметров алгоритма по сравнению со средним квадратичным отклонением.
Пусть v – изображение на выходе сети, которое по размерам и количеству каналов совпадает с y, тогда функция потери выглядит следующим образом:

$$H\left(y\left(n_1,n_2\right),v\left(n_1,n_2\right)\right)=-\sum_{i=0}^{m-1}y\left(n_1,n_2,i\right)\log v(n_1,n_2,i) \tag{3}$$

Функционал качества сети на обучающей выборке с использованием функции потери равен:

$$Q(X^G)=-\frac{1}{G}\sum_{i=0}^{G-1}\sum_{j=0}^{N_1-1}\sum_{k=0}^{N_2-1}\sum_{t=0}^{m-1}y_i\left(j,k,t\right)\log\left(v_i\left(j,k,t\right)\right) \tag{4}$$

В качестве алгоритма обучения выступал метод адаптивного стохастического градиента Adam [10]. Во время этапа настройки параметров сети использовалась техника уменьшения коэффициента обучения в случае, если значение фукционала качества сети на валидационной выборке не улучшалось.

4. Результаты экспериментов

Обучение и тестирование модели проводилось на изображениях, полученных с помощью программы Unity. Количество камер K было равным 8, количество сцен S – 23, а количество RGB изображений в элементе выборки m – 5. Размер изображений подаваемых на вход составлял 300×300.

Таким образом, на вход подавалось изображение с 15 каналами. Первые 3 канала принадлежали опорному кадру. Пример опорного кадра, не опорного кадра, требуемый ответ и выход нейронной сети представлены на рисунках 4, 5, 6, 7. Яркость на рисунках 6 и 7 отвечает за количество сопряженных пар, которые появляются при сопоставлении опорного кадра и изображений с других камер.

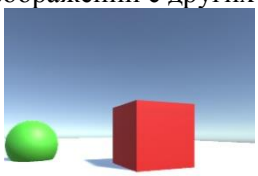

Рисунок 4. Опорный кадр.

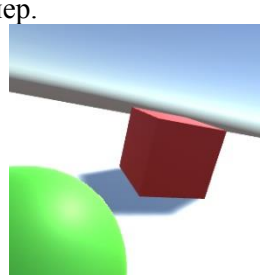

Рисунок 5. Не опорный кадр.

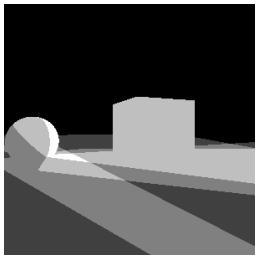

Рисунок 6. Требуемый ответ для сети.

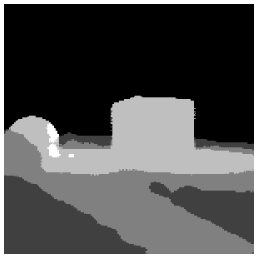

Рисунок 7. Выход сети.

В качестве метрики для оценки результата использовалась точность оценки количества сопряженных пар:

$$M = \frac{1}{N_1 N_2 O} \sum_{t=0}^{O-1} \sum_{i=0}^{N_1-1} \sum_{j=0}^{N_2-1} \left(\arg\max_u v_t(u,i,j) == \arg\max_l y_t(l,i,j) \right), \quad (5)$$

где O – количество элементов в тестовой выборке, а переменные u и l пробегают по каналам выходного изображения нейронной сети и требуемого на выходе изображения соответственно. В роли тестовой выборки выступали изображения, полученные со сцен, не входившие в обучающую выборку.

После обучения и проверки работы сети точность поиска количества сопряжённых пар составила 0.96. Предложенный метод оценки количества сопряжнных пар с компактной архитектурой нейронной сети даёт возможность использовать алгоритм в режиме реального времени на мобильных графических процессорах.

5. Заключение

В данной статье представлен метод попиксельной оценки восстановления трехмерных сцен по последовательности изображений с использованием свёрточных нейронных сетей. Была описана архитектура свёрточной сети и предложена процедура обучения.

Проведена серия экспериментов, в ходе которой выяснено, что разработанный метод демонстрирует хорошую эффективность в оценке восстановления трехмерных сцен путем оценки количества сопряжённых пар. Дальнейшие исследования будут направлены на разработку метода оценки вектора перемещения камеры в пространстве.

6. Литература

[1] Horn, B.K.P. Robot vision / B.K.P Horn. – MIT Press, 1986. – 614 p.

[2] Choy, B.C. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction / B.C. Choy, D. Xu, J.Y. Gwak, K. Chen // European Conference on Computer Vision. – 2016. – Vol. 394(2). – P. 309-345.
[3] Lowe, D. Distinctive image features from scale-invariant keypoints / D. Lowe // International Journal of Computer Vision. – 2004. – Vol. 130(2). – P. 91-110.
[4] Zbontar, J. Computing the stereo matching cost with a convolutional neural network / J. Zbontar, Y. LeCun // Conference on Computer Vision and Pattern Recognition. – 2015. – Vol. 324(1). – P. 234-246.
[5] Shelhamer, E. Fully convolutional networks for semantic segmentation / E. Shelhamer, J. Long, T. Darrell // The Pattern Analysis and Machine Intelligence. – 2016. – Vol. 324(5). – P. 100-108.
[6] Ronneberger, O. U-net: Convolutional networks for biomedical image segmentation / O. Ronneberger, P. Fischer, T. Brox // Medical Image Computing and Computer-Assisted. – 2015. – Vol. 345(1). – P. 234-241.
[7] Badrinarayanan, V. Segnet: A deep convolutional encoder-decoder architecture for image segmentation / V. Badrinarayanan, A. Kendall, R. Cipolla // IEEE Conferenceon Computer Vision and Pattern Recognition. – 2015. – Vol. 423(4). – P. 125-145.
[8] Chaurasia, A. Linknet: Exploiting encoder representations for efficient semantic segmentation / A. Chaurasia, E. Culurciello // IEEE Conference on Computer Visionand Pattern Recognition. – 2017. – Vol. 362(3). – P. 234-247.
[9] Golik, P. Cross-entropy vs. squared error training: a theoretical and experimental comparison / P. Golik, P. Doetsch, H. Ney // The 14th Annual Conference of the International Speech Communication Association. – 2019. – Vol. 234(4). – P. 100-105.
[10] Diederik, P. Adam: A Method for Stochastic Optimization / P. Deiderik // The Pattern Analysis and Machine Intelligence. – 2014. – Vol. 723(3) – P. 624-637.

Благодарности
Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 18-01-00748-а, № 17-29-03190-офи-м.

Possibility estimation of 3D scene reconstruction from multiple images

E.A. Dmitriev[1], V.V. Myasnikov[1,2]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Image Processing Systems Institute of RAS - Branch of the FSRC "Crystallography and Photonics" RAS, Molodogvardejskaya street 151, Samara, Russia, 443001

Abstract. This paper presents a pixel-by-pixel possibility estimation of 3D scene reconstruction from multiple images. The method is based on conjugate pairs number estimation with convolutional neural networks. The efficiency criteria of algorithm are the resulting estimation accuracy and speed of the algorithm. All experiments were conducted on images that were got with Unity program. The results of experiments showed effectiveness of our approach in 3D scene reconstruction problem.