Глубокие сверточные автоэнкодеры: стереоотождествление для восстановления трехмерных моделей слабо текстурированных объектов*

***В. В. Князь*[1,2], *О. В. Выголов*[1], *В. В. Федоренко*[1], *В. Д. Севрюков*[1]**
vl.kniaz@gosniias.ru; o.vygolov@gosniias.ru; vfedorenko@gosniias.ru; vsevryukov@gosniias.ru
[1]ФГУП «Государственный научно-исследовательский институт авиационных систем», Россия, г. Москва, ул. Викторенко, 7
[2]Московский физико-технический институт, Россия, г. Долгопрудный, Институтский пер., 9

Восстановление трехмерных (3D) моделей объектов со слабо выраженными текстурами требует использования дескрипторов, способных разделять очень похожие друг на друга классы характерных точек. К таким объектам, например, относятся артефакты, найденные в ходе археологических раскопок, покрытые равномерным слоем грунта. Широко распространенные дескрипторы особых точек (SIFT — scale-invariant feature transform, SURF — speeded up robust features) часто не справляются с задачей стереоотождествления в случае слабо выраженных текстур. Рассматривается новый метод решения данной задачи на основе глубоких сверточных автоэнкодеров (САЭ). Автоэнкодер (АЭ) производит понижение размерности изображения на несколько порядков и формирует код, который может использоваться для решения задачи стереоотождествления. Рассмотрена архитектура АЭ, производящего кодирование и восстановление цветных изображений, разрешением 32×32 пиксела. Приводится сравнение результатов работы предложенного метода стереоотождествления и классических дескрипторов особых точек. Экспериментально восстановлены 3D модели археологических раскопок, производимых в ходе Босфорской экспедиции, организованной Государственным историческим музеем. Анализ полученных результатов показывает, что предложенный метод превосходит существующие дескрипторы особых точек на слабо текстурированных объектах и позволяет успешно решать задачу стереоотождествления для восстановления 3D моделей.

Ключевые слова: *глубокие сверточные нейронные сети; автоэнкодеры; стереоотождествление*

DOI: 10.21469/22233792.3.2.03

1 Введение

Поиск соответствующих точек на изображениях является ключевым этапом процесса восстановления 3D моделей объектов по снимкам, полученным с камеры, пространственное и угловое положение (внешнее ориентирование) которой является неизвестным. Классические дескрипторы особых точек, подобные дескрипторам SIFT [1] или SURF [2], позволяют решать задачу стереоотождествления для снимков объектов с характерными текстурами, сделанных с высоким разрешением. Решение такой задачи для объектов со слабо выраженными или отсутствующими текстурами до недавнего времени являлось невозможным.

*Работа выполнена при финансовой поддержке РФФИ, проекты № 17-29-04410 и № 16-08-01260.

Появление технологий глубокого обучения коренным образом изменило состояние дел на передовом фронте машинного зрения. В течение нескольких лет ряд сложных задач, недоступных классическим алгоритмам машинного зрения, был успешно решен с помощью глубоких сверточных нейронных сетей (ГСНС).

В данной статье рассматривается архитектура ГСНС для стереоотождествления объектов со слабо выраженными текстурами. В основе архитектуры лежит САЭ, который представляет собой нейронную сеть, воспроизводящую на выходе входное изображение. Данная сеть состоит из двух частей: энкодера, осуществляющего сжатие входного изображения в код из нескольких элементов, и декодера, реализующего восстановление исходного изображения из кода. Таким образом, АЭ понижает размерность изображения на несколько порядков и формирует код, который может использоваться для решения задачи стереоотождествления.

Рассмотрена архитектура АЭ, производящего кодирование и восстановление цветных изображений, разрешением 32×32 пиксела. Сравниваются результаты работы предложенного метода стереоотождествления и классических дескрипторов особых точек. Экспериментально восстановлены 3D модели археологических раскопок, производимых в ходе Босфорской экспедиции, организованной Государственным историческим музеем. Анализ полученных результатов показывает, что предложенный метод превосходит существующие дескрипторы особых точек на слабо текстурированных объектах и позволяет успешно решать задачу стереоотождествления для восстановления 3D моделей.

2 Обзор работ в данной области

Первые методы восстановления 3D объектов по изображениям появились более 50 лет назад. К наиболее интенсивно развивающимся областям компьютерного зрения в этом случае можно отнести: восстановление 3D-моделей по снимкам монокулярной камеры [3, 4], оценку положения камеры по одному снимку (6 degrees of freedom (6DoF)) pose estimation) [5] и автоматический анализ окружающей сцены. Надежное стереоотождествление является ключевым элементом большинства предлагаемых подходов. Большинство современных алгоритмов, осуществляющих восстановление 3D-моделей, используют подходы, основанные на аналитически разработанных дескрипторах характерных точек [1, 2].

Постоянное повышение разрешения камер позволило создать надежные методы стереоотождествления для объектов с характерными текстурами [1, 2, 6]. Стандартные подходы, такие как метод «структура сцены из движения» (Structure from Motion, SfM) [7], производят оценку параметров внутреннего и внешнего ориентирования камеры на основе поиска соответствующих характерных точек. При этом, если на объекте отсутствуют контрастные текстуры, качество оценки ориентирования камеры значительно падает. Альтернативный подход предполагает использование кодированных 3D контрольных точек с известными 3D координатами, которые могут быть однозначно идентифицированы на всех снимках. Как показано в работе [8], калибровка камеры с использованием кодированных точек может быть осуществлена с высокой точностью и использована для последующего восстановления 3D-моделей объектов согласно классическому фотограмметрическому процессу.

В последние годы появились методы, не использующие поиск характерных точек для восстановления плотной 3D модели сцены, такие как LSD-SLAM (Large-Scale Direct monocular Simultaneous Localization And Mapping) [4]. Тестирование данных методов на слаботекстурированных объектах [9] показало, что данные методы обеспечивают низкое качество 3D-модели в случае низкой контрастности текстур на исходных изображениях.

Методы стереоотождествления, которые используют разбиение на конечные дискретные плоскости, такие как метод сметающих плоскостей (plane sweep mathcing) или PatchMatch [10–12], обеспечивают надежную работу на слаботекстурированных объектах. Однако для восстановления 3D-моделей объектов с использованием данных методов необходимо, чтобы восстанавливаемая поверхность обладала диффузными или ламбертовскими отражающими свойствами и была локально гладкой.

Появившиеся в последние годы дескрипторы, основанные на методах глубокого обучения [5, 13, 14], превосходят своих предшественников по скорости работы и точности поиска соответствующих точек. Дескрипторы, основанные на глубоком обучении, можно разделить на две группы.

Первая группа основана на классических глубоких нейронных сетях для классификации изображений [15, 16]. Для решения задачи стереоотождествления верхние слои сети удаляются. Значения, полученные на выходе оставшихся слоев, используются для поиска соответствующих точек.

Вторая группа дескрипторов, основанных на глубоком обучении, использует методы обучения «без учителя». Это обусловлено тем, что множество всех возможных характерных точек на изображениях не ограничено и невозможно выбрать хорошие классы точек на этапе обучения. В [5] предлагается использовать САЭ для решения данной проблемы. Использование САЭ для оценки внешнего ориентирования камеры по одному изображению видимого диапазона и соответствующей карте глубины (RGB-D) показало превосходные результаты при тестировании на различных наборах данных. Другим преимуществом САЭ является устойчивая работа при поступлении данных, значительно отличающихся от состава обучающей выборки. Таким образом, методы, основанные на глубоком обучении, предоставляют гибкий подход к решению задачи стереоотождествления, который обеспечивает надежную работу для объектов со слабо выраженными текстурами.

Для получения окончательной 3D-модели объекта требуется произвести сгущение облака точек. Большинство известных алгоритмов сгущения облаков точек используют характерные точки, основанные на яркости пикселей. Метод поиска характерных точек, основанный на выделении силуэтов [17], и метод 3D разреза графов [18] требуют наличия острых граней на объектах для надежной работы. Следовательно, они не могут быть непосредственно использованы для сгущения облаков точек на объектах со слабо выраженными текстурами. Альтернативным подходом является алгоритм полуглобального отождествления (Semiglobal Matching, SGM [19]), показавшим себя надежным решением для сгущения облака точек по снимкам объектов со слабо выраженными текстурами.

3 Стереоотождествление с использованием автоэнкодеров

В этом разделе последовательно рассматриваются основные этапы предлагаемого метода восстановления 3D-моделей. Представлены метод поиска характерных точек, архитектура используемого САЭ, целевая функция и обучающая выборка. Приведено описание модифицированного метода SfM, используемого для восстановления 3D-модели объекта.

3.1 Поиск характерных точек

Классический подход к выделению характерных точек на изображении основывается на детекторах особых точек. Тестирование классических детекторов характерных точек на объектах со слабо выраженными текстурами [20] показало, что подобные детекторы не гарантируют надежного выделения постоянного множества характерных точек. Поскольку к разрабатываемому методу не выдвигаются требования по быстродействию, наиболее

Рис. 1 Пример выбора характерных точек на изображении раскопа

целесообразным представляется выделение особых точек с помощью скользящего окна с равномерной сеткой. Пример полученных характерных точек представлен на рис. 1. Выбор размера скользящего окна зависит от исходного разрешения выбранной камеры. Учитывая архитектуру, предложенную в работе [5], был выбран размер окна 32×32 пиксела.

3.2 Сверточный автоэнкодер

Автоэнкодер является частным случаем нейронной сети с опережающей обратной связью. Автоэнкодер принимает на вход сигнал $\boldsymbol{x}$ и пытается воспроизвести его точную копию на выходе $\boldsymbol{y}$ [21]. Сеть состоит из двух основных частей. Первая часть — это кодер $\boldsymbol{h} = f(\boldsymbol{x})$, который сжимает входные данные $\boldsymbol{x}$ с помощью скрытого слоя $\boldsymbol{h}$, который создает код F, содержащий все значения, необходимые для восстановления входного сигнала. Вторая часть — это декодер, который пытается восстановить входные данные $\hat{\boldsymbol{y}} = g(F)$. Поскольку код F имеет значительно более низкую размерность, чем исходное изображение, во время обучения АЭ пытается создать функцию, которая фиксирует наиболее важные черты изображений в обучающей выборке. После этапа обучения выход скрытого слоя можно использовать в качестве кода F, который может использоваться для эффективного поиска соответствующего изображения. Недавние исследования [5, 22] показали, что конволюционные слои повышают качество реконструкции АЭ. Автоэнкодеры с конволюционными и деконволюционными слоями обычно называют сверточными.

К разрабатываемой эффективной архитектуре САЭ для решения задачи стереоотождествления были выдвинуты следующие требования. Во-первых, САЭ должен принимать на вход изображения низкого разрешения. Во-вторых, он должен иметь небольшое количество обучаемых параметров для достижения хороших свойств сходимости на малых обучающих выборках.

Изначально в качестве исходной архитектуры была выбрана модель, предложенная в [22]. Данный САЭ был реализован для библиотеки Caffe. Обучение на выборке MNIST [23] продемонстрировало превосходные результаты. Однако обучение на более сложной выборке CIFAR-10 [24] показало, что данный САЭ имеет тенденцию сходиться к среднему значению всех изображений в выборке. После анализа результатов обучения был сделан вывод, что архитектура [22] имеет недостаточное количество обучаемых параметров. Для решения этой проблемы было увеличено количество обучаемых сверточных фильтров. В качестве опорной архитектуры, была взята модель, предложенная

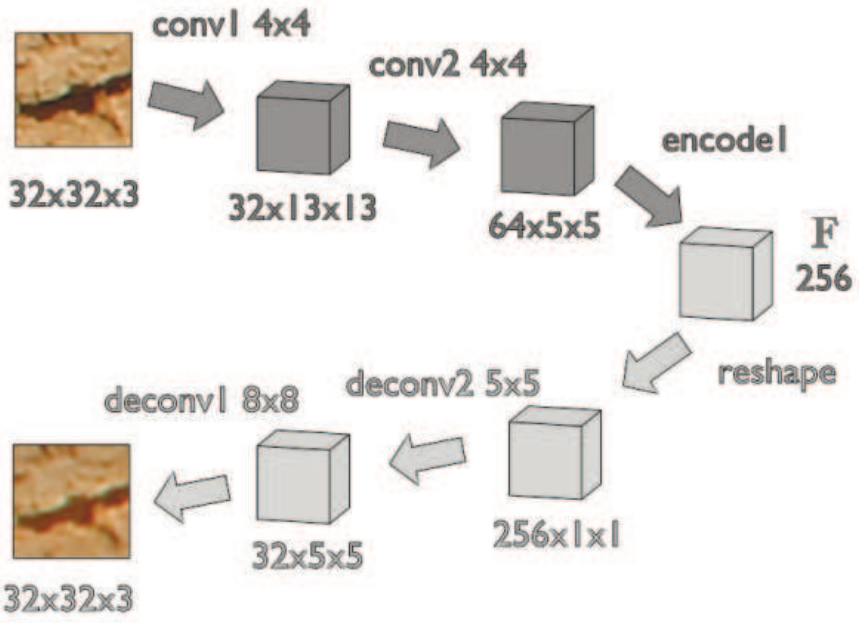


Рис. 2 Архитектура сети САЭ

Таблица 1 Архитектура сети САЭ (F — размер обучаемого кода)

Слой	Размер на выходе слоя	Фильтр	Шаг (stride)
Входной сигнал	$3 \times 32 \times 32$		
Конволюционный	$8 \times 13 \times 13$	4×4	2
Конволюционный	$16 \times 5 \times 5$	4×4	2
Полносвязный	F		
Деконволюционный	$8 \times 5 \times 5$	5×5	2
Деконволюционный	$1 \times 32 \times 32$	8×8	5

в [25]. Для решения поставленной задачи в архитектуре были произведены следующие изменения. Во-первых, изменены размеры конволюционных и деконволюционных слоев для достижения целевого размера изображения в 32×32 пиксела. Во-вторых, слои активации tanh были заменены на слои sigmoid для достижения более эффективного обратного распространения ошибок [26] и повышения стабильности обучения на малых обучающих выборках. Для определения рационального размера кода F, обеспечивающего разумный компромисс между качеством восстановления и степенью сжатия, были проведены три эксперимента с различными размерами кода: 64, 128 и 256 байт. Нами был выбран размер кода в 256 байт. Окончательная архитектура сети САЭ представлена на рис. 2 и в табл. 1.

Были опробованы две функции потерь. Одна из них — классическая, основанная на евклидовом расстоянии, — не обеспечивает равномерной сходимости для архитектур ГСНС с деконволюционными слоями [27], поэтому используется кросс-энтропийная потеря (логистическая функция потерь), задаваемая выражением:

$$L_{\text{sc}} = -\frac{1}{n} \sum_{i \in (w,h)} \Big(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \Big),$$

где w и h — размеры слоя; y — значение пикселя целевого изображения в точке (x, y); $\hat{y}$ — значение пикселя изображения, восстановленного САЭ, в точке (x, y).

Было установлено, что добавление гауссовского шума непосредственно перед заключительным слоем активации повышает стабильность процесса обучения [21]. Стандартное отклонение гауссовского шума равномерно увеличивается от 0 до 0,5 во время процесса обучения.

3.3 Обучающая выборка

Для успешного обучения САЭ требуется большая обучающая выборка, обеспечивающая высокое разнообразие фрагментов изображений. Для обучения разработанной архитектуры САЭ была сформирована обучающая выборка объемом в 1 млн фрагментов изображений. Для формирования обучающей выборки использовались технологии 3D моделирования. С использованием фотограмметрического сканирующего комплекса на основе структурированного подсвета [28] были получены текстурированные 3D модели тестового раскопа. Полученные модели были импортированы в среду 3D моделирования Blender. С использованием Blender API на языке Python были разработаны автоматизированные сценарии построения фрагментов изображения заданных 3D точек с различных ракурсов.

Для создания обучающей выборки была использована методика, аналогичная [29]. Виртуальная камера размещается на поверхности икосаэдра и направляется на характерные точки. Последние были выбраны с использованием детектора углов Харриса [30] на исходных изображениях. Трехмерные координаты точек были получены с помощью обратной проекции в 3D пространство.

Для определения соответствующих точек каждой характерной точке с известными 3D координатами назначается уникальный идентификатор. Идентификатор точки сохраняется в таблице соответствий, сопоставляющей код F, выдаваемый САЭ, с идентификатором. Для каждого идентификатора было сформировано 7000 фрагментов изображений, показывающих точку с разных ракурсов.

3.4 Поиск соответствующих точек

Для создания таблицы соответствий используются все фрагменты изображений из обучающей выборки. Стереоотождествление осуществляется с использованием метода голосования. Для заданного входного фрагмента изображения $\boldsymbol{I}$ формируется код F с помощью САЭ. После этого из таблицы соответствия запрашивается n ближайших соседей. Идентификатор входного фрагмента $d(\boldsymbol{I})$ задается идентификатором большинства голосов найденных соседей. Чтобы отфильтровать ложные гипотезы, вычисляется вероятность p того, что фрагмент $\boldsymbol{I}$ имеет идентификатор z. Вероятность p соответствует отношению количества соседей с идентификатором z к выбранному числу ближайших соседей n:

$$p = P(d(\boldsymbol{I}) = z) = \frac{|\{b \in B : b = z\}|}{n},$$

где B — множество идентификаторов ближайших соседей.

Пусть $\boldsymbol{I}_1$, $\boldsymbol{I}_2$ — два изображения, которые требуется сопоставить. Тогда вероятность p_{pair} того, что у данных изображений одинаковый идентификатор, определяется следующим образом:

$$p_{\text{pair}} = \begin{cases} P(d(\boldsymbol{I}_1) = z)P(d(\boldsymbol{I}_2) = z), & d(\boldsymbol{I}_1) = d(\boldsymbol{I}_2)\,; \\ 0 & \text{в противном случае.} \end{cases}$$

Использование голосования для поиска соответствующих точек продемонстрировало высокие результаты для решения задачи оценки положения заданного объекта [5]. Для использования этого подхода в задаче стереоотождествления требуется построение обширной таблицы соответствий, в которой содержатся фрагменты изображений, близкие по виду к фрагментам изображений объекта интереса. Для оценки работоспособности разработанного метода при тестировании использовалась таблица соответствий, построенная по 3D-модели археологического раскопа. Следует отметить, что тестовая и обучающая выборки не пересекались.

4 Калибровка камеры и оценка внешнего ориентирования

В данном разделе рассматривается метод калибровки камеры. Наилучшие кандидаты среди полученных соответствующих точек отбираются с помощью пороговой фильтрации и используются для оценки внешнего ориентирования.

4.1 Калибровка камеры

Для построения высокоточной 3D модели объекта требуется предварительный этап калибровки камеры. Калибровка камеры осуществляется согласно модели, предложенной в [31], с помощью специализированного оригинального программного обеспечения [8]. Для выполнения калибровки используется тестовое поле с известными пространственными координатами опорных точек (рис. 3).

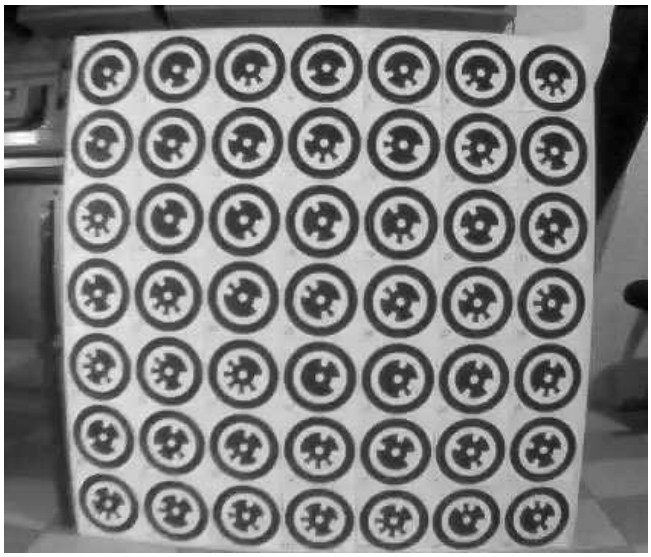


Рис. 3 Тестовое поле для калибровки камеры

Для калибровки было получено 20 изображений. Среднеквадратичная ошибка в контрольных точках для оцениваемых параметров модели камеры составила около 0,1 мм. Полученные параметры внутреннего ориентирования камеры использовались для оценки положения камеры и восстановления 3D-моделей.

4.2 Оценка внешнего ориентирования камеры

Для определения внешнего ориентирования камеры используются соответствующие точки, положения которых были найдены с использованием САЭ. Сверточный АЭ позволяет найти соответствующие точки на двух изображениях, но точность найденных координат соответствующих точек недостаточна для оценки внешнего ориентирования. Для

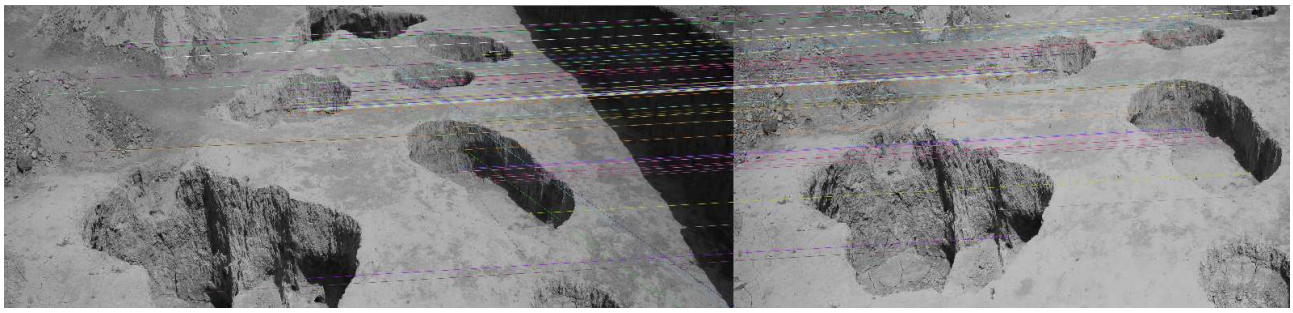


Рис. 4 Поиск соответствующих точек

уточнения начальной оценки координат соответствующих точек используется субпиксельная корреляция. Пример поиска соответствующих точек приведен на рис. 4.

Для оценки положения камеры использовали нелинейную минимизацию невязок измерения (обратной проекции) с помощью блочного уравнивания [32, 33]. Для минимизации квадратичной ошибки обратной проекции обнаруженных точек используется избыточное число уравнений, содержащих координаты найденных точек $\boldsymbol{x}_{ij}$ на изображении j, представленных как функции от неизвестных параметров внешнего ориентирования $(\boldsymbol{R}, \boldsymbol{X})$ и неизвестных 3D координат точки $\boldsymbol{p}_i$. Решение системы уравнений осуществляется с помощью метода наименьших квадратов. Система уравнений имеет вид:

$$\boldsymbol{x}_{ij} = \boldsymbol{f}(\boldsymbol{p}_i, \boldsymbol{R}_j, \boldsymbol{X}_j). \tag{1}$$

Ошибка обратной проекции задается выражением:

$$\boldsymbol{E} = \sum_{i,j} \left(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \Delta \boldsymbol{x} + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{R}} \Delta \boldsymbol{R} + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{X}} \Delta \boldsymbol{X} - \boldsymbol{r}_{ij} \right), \tag{2}$$

где $\boldsymbol{r}_{ij} = \boldsymbol{x}_{ij} - \hat{\boldsymbol{x}}_{ij}$ — вектор текущей невязки (двумерная ошибка в оценке положении точки); частные производные заданы относительно неизвестных параметров внешнего ориентирования (пространственное и угловое положение камеры).

Для определения реального масштаба объекта используется известное расстояние между опорными точками, заданными в рабочей области.

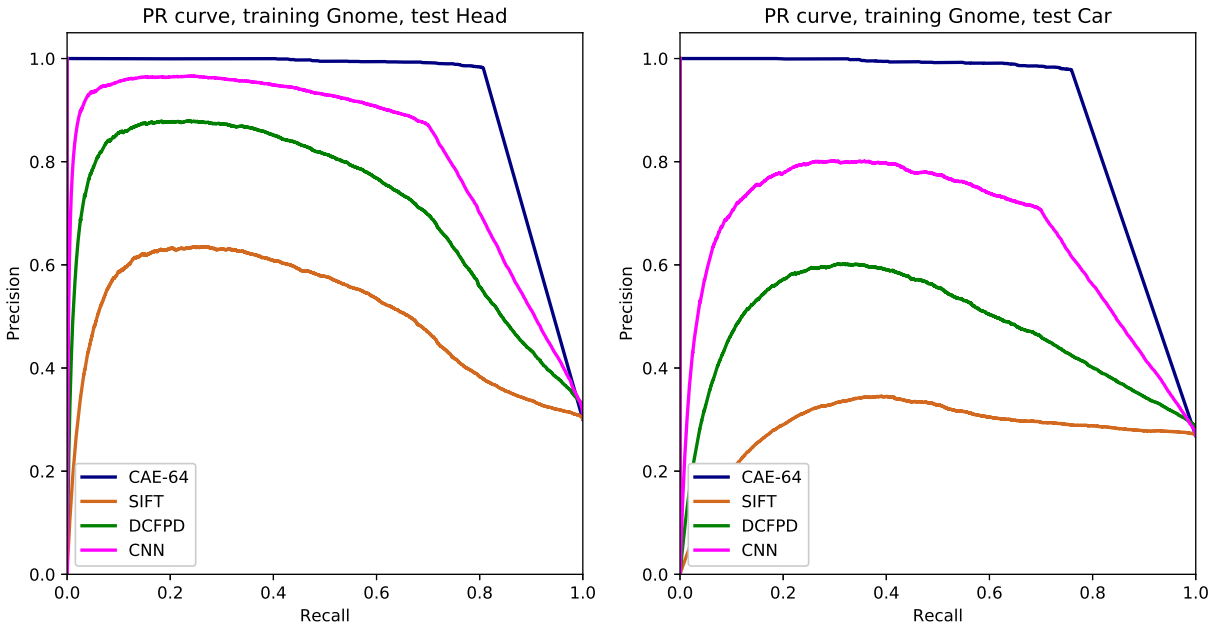
Точность оценки параметров внешнего ориентирования зависит от конфигурации использованных точек на исходных обрабатываемых изображениях. Тестирование оценки внешнего ориентирования осуществлялось с помощью набора кодированных меток, расставленных на модели раскопа. Точные 3D координаты меток были получены с использованием лазерного дальномера, обеспечивающего точность измерения 2 мм. Относительная ошибка в оценке положения камеры, полученном с помощью блочного уравнивания и точно найденного по набору кодированных меток, лежит в пределах 2%. Таким образом, предложенный метод обеспечивает достаточную точность оценки положения камеры.

5 Сравнение результатов работы предложенного метода стереоотождествления и классических дескрипторов особых точек

Произведем сравнение, дискриминирующее качество кодов, создаваемых с помощью САЭ и SIFT (табл. 2).

Таблица 2 Сравнение колов SIFT и САЭ-64

Выборка	SIFT	САЭ-64
Гном	0,51	0,93
Гном	0,28	0,89



(*а*) Фотография (*б*) Восстановленная 3D-модель

Рис. 5 Сравнение различных методов

Для оценки качества метода используется PR-кривая и область под кривой (area under curve, AUC) в качестве показателей производительности. Для тестирования использовалось 10 000 положительных пар фрагментов изображений и 30 000 отрицательных пар. Пары сравниваются с использованием евклидова расстояния между кодами F_1 и F_2 для двух фрагментов изображений. Для тестирования использовалась реализация SIFT из проекта VLFeat (рис. 5).

6 Восстановление трехмерных моделей археологических раскопок

Анализ работоспособности предложенного метода восстановления 3D моделей слабо текстурированных объектов производился с использованием данных, полученных Босфорской экспедицией, организованной Государственным историческим музеем. Целью данной экспедиции является исследование останков поселений древнегреческой колонии, располагавшейся на территории современного Краснодарском края недалеко от станицы Голубицкая. Неотъемлемой частью археологических раскопок является документирование полученных раскопов и найденых артефактов. Перспективным способом документирования хода раскопок является получение 3D моделей с помощью фотограмметрических методов. Классический фотограмметрический подход для слабо текстурированных объектов предполагает предварительную расстановку кодированных меток или использование

(*а*) Фотография (*б*) Восстановленная 3D-модель

Рис. 6 Восстановление 3D модели раскопа

структурированного подсвета. Недостатком данного подхода является большая длительность процесса документирования, замедляющая общий темп раскопок.

Предложенный метод на основе САЭ устраняет данный недостаток и позволяет свести документирование хода раскопок к съемке набора цветных фотографий высокого разрешения. Для тестирования работы предложенного метода были осуществлены съемки двух раскопов, площадью около 100 м^2 каждый. Съемка осуществлялась с использованием фотоаппарата Canon EOS M с фокусным расстоянием объектива 20 мм. Разрешение исходных снимков составило 5184×3456 пиксел.

Для поиска соответствующих точек использовались таблица соответствий, построенная с использованием 3D модели тестового раскопа. На первом этапе были выбраны все фрагменты изображения с высоким значением среднеквадратического отклонения. Для каждого фрагмента были найдены идентификатор типового фрагмента изображения тестового раскопа и вероятность p. После этого были отобраны все те фрагменты, для которых значение вероятности $p > 0{,}9$. С использованием фрагментов изображений с высокой вероятностью p были найдены пары снимков с большим коэффициентом перекрытия. Окончательная оценка положений камер и расчет 3D облака точек были произведены методом блочного уравнивания с использованием формул (1) и (2). Вид полученной 3D модели с наложенными текстурами приведен на рис. 6. Анализ полученной 3D модели показал работоспособность предложенного метода. Сравнение измерений, сделанных по модели, с опорными измерениями, сделанными на местности с использованием лазерного дальномера показало, что ошибка 3D модели лежит в пределах 5%.

Для оценки качества работы метода для произвольного объекта была восстановлена 3D модель древнегреческой амфоры, найденной в ходе раскопок. Качество облака точек, полученного с использованием таблицы соответствий для тестового раскопа, было недостаточным для триангуляции и построения текстурированной модели. Данный результат связан с тем, что характерные текстуры, присутствующие на амфоре, не были представлены в таблице соответствий для тестового раскопа. Для восстановления детальной 3D модели амфоры была построена новая таблица соответствий с использованием 3D модели амфоры, взятой из Интернета. Использование новой таблицы соответствий позволило заметно повысить плотность облака точек и получить полную 3D модель амфоры (рис. 7).

(*а*) Фотография (*б*) Восстановленная 3D-модель

Рис. 7 Восстановление 3D модели амфоры, найденной в ходе раскопок

7 Заключение

В работе представлен новый метод поиска соответствующих точек на изображениях на основе глубокого обучения. Предложена архитектура САЭ, принимающего на вход цветное изображение разрешением 32×32 пиксела и формирующего код, позволяющий находить похожие друг на друга фрагменты изображений. Для осуществления поиска соответствующих фрагментов производится обучение автоэнкодера на фрагментах изображений, похожих на фрагменты изображения целевого объекта. По обучающей выборке строится таблица соответствий, которая определяет набор классов характерных фрагментов изображений (углы, линии и т. д.) Для поиска соответствующих фрагментов изображений для каждого фрагмента строится код с помощью АЭ. По коду производится запрос в таблице соответствий для определения класса заданного фрагмента. Стереоотождествление осуществляется на основе метода голосования.

Предложенная архитектура была реализована с использованием библиотеки Caffe. Обучение произведено с использованием графического процессора NVidia Titan X. Тестирование архитектуры показало, что она хорошо обучается на фрагментах изображений произвольных классов объектов.

Произведено тестирование предложенного метода для восстановления 3D моделей раскопок, проводимых в ходе Босфорской экспедиции Государственного исторического музея. Анализ точности восстановленных моделей показал, что разработанный метод обеспечивает точность, необходимую для документирования хода раскопок.

Литература

[1] *Lowe D. G.* Object recognition from local scale-invariant features // Conference (International) on Computer Vision Proceedings. — Washington, DC, USA: IEEE Computer Society, 1999. P. 1150. http://dl.acm.org/citation.cfm?id=850924.851523.

[2] *Bay H., Tuytelaars T., Van Gool L.* Surf: Speeded up robust features // ECCV, 2006. P. 404–417.

[3] *Klein G., Murray D. W.* Parallel tracking and mapping for small AR workspaces // ISMAR, 2007. P. 225–234.

[4] *Engel J., Schöps Th., Cremers D.* LSD-SLAM: Large-scale direct monocular SLAM // ECCV, 2014. Vol. 8690. Ch. 54. P. 834–849.

[5] *Kehl W., Milletari F., Tombari F., Ilic S., Navab N.* Deep learning of local RGB-D patches for 3D object detection and 6D pose estimation // ECCV, 2016. Vol. 9907. No. 7. P. 205–220.

[6] *Rublee E., Rabaud V., Konolige K., Bradski G.* ORB: An efficient alternative to SIFT or SURF // Conference (International) on Computer Vision, 2011. P. 2564–2571.

[7] *Remondino F., Spera M. G., Nocerino E., Menna F., Nex F.* State of the art in high density image matching // Photogramm. Rec., 2014. Vol. 29. No. 146. P. 144–166.

[8] *Knyaz V. A., Chibunichev A. G.* Photogrammetric techniques for road surface analysis // Int. Arch. Photogramm., 2016. Vol. XLI-B5. P. 515–520. doi: 10.5194/isprs-archives-XLI-B5-515-2016. https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLI-B5/515/2016/.

[9] *Yamaguchi M., Saito H., Yachida S.* Application of LSD-SLAM for visualization temperature in wide-area environment // VISIGRAPP, 2017. P. 216–223.

[10] *Gallup D., Frahm J.-M., Mordohai Ph., Yang Q., Pollefeys M.* Real-time plane-sweeping stereo with multiple sweeping directions // CVPR, 2007. P. 1–8.

[11] *Bleyer M., Rhemann Ch., Rother C.* PatchMatch stereo — stereo matching with slanted support windows // British Machine Vision Conference, 2011. P. 14.1–14.11.

[12] *Galliani S., Lasinger K., Schindler K.* Gipuma: Massively parallel multi-view stereo reconstruction // Publ. Deutschen Gesellschaft Photogrammetrie Fernerkundung Geoinformation e. V, 2016. Vol. 25. P. 361–369.

[13] *Wohlhart P., Lepetit V.* Learning descriptors for object recognition and 3D pose estimation. arXiv.org, 2015. arXiv:1502.05908.

[14] *Simo-Serra E., Trulls E. Ferraz L., Kokkinos I., Fua P., Moreno-Noguer F.* Discriminative learning of deep convolutional feature point descriptors // ICCV, 2015. P. 118–126.

[15] *Krizhevsky A., Sutskever I., Hinton G. E.* Imagenet classification with deep convolutional neural networks // Advances in Neural Information Processing Systems, 2012. P. 1095–1105.

[16] *Szegedy Ch., Liu W., Jia Y., et al.* Going deeper with convolutions // IEEE Conference on Computer Vision and Pattern Recognition. — IEEE, 2015. P. 1–9.

[17] *Esteban C. H., Schmitt F.* Silhouette and stereo fusion for 3D object modeling // 4th Conference (International) on 3D Digital Imaging and Modelling Proceedings. — Banff, Alberta, Canada, 2003. P. 46–53.

[18] *Vogiatzis G., Hernandez C., Torr P. H. S., Cipolla R.* Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency // IEEE T. Pattern Anal., 2007. Vol. 29. No. 12. P. 2241–2246.

[19] *Bethmann F., Luhmann T.* Semi-global matching in object space // Int. Arch. Photogramm., 2015. Vol. XL-3/W2. P. 23–30.

[20] *Hajebi K., Zelek J. S.* Structure from infrared stereo images // Canadian Conference on Computer and Robot Vision Proceedings, 2008. P. 105–112.

[21] *Goodfellow I., Bengio Y., Courville A.* Deep learning. — The MIT Press, 2016. 800 p.

[22] *Turchenko V., Luczak A.* Creation of a deep convolutional auto-encoder in caffe // CoRR abs/1501.02565, 2015. Vol. 1512. arXiv:1512.01596.

[23] *LeCun Y., Bottou L., Bengio Y., Haffner P.* Gradient-based learning applied to document recognition // Proc. IEEE, 1998. Vol. 86. Iss. 11. P. 2278–2324.

[24] *Krizhevsky A., Hinton G.* Learning multiple layers of features from tiny images. — University of Toronto, 2009. Technical Report.

[25] *Ridgeway K., Snell J., Roads B., Zemel R. S., Mozer M. C.* Learning to generate images with perceptual similarity metrics // CoRR, 2015. arXiv:1511.06409. http://arxiv.org/abs/1511.06409.

[26] *LeCun Y., Bottou L., Orr G. B, Müller K. R.* Efficient BackProp // Neural networks: Tricks of the trade / Eds. G. B. Orr, K.-R. Müller. — Lecture notes in computer science ser. — Berlin–Heidelberg: Springer, 1998. Vol. 1524. P. 9–50.

[27] *Zhang R., Isola P., Efros A. A.* Colorful image colorization // ECCV, 2016. Vol. 9907. Ch. 40. P. 649–666.

[28] *Knyaz V. A.* Multi-media projector — single camera photogrammetric system for fast 3D reconstruction // Int. Arch. Photogramm., 2010. Vol. XXXVIII-5. P. 343–348. http://www.isprs.org/proceedings/XXXVIII/part5/papers/143.pdf.

[29] *Hinterstoisser S., Lepetit V., Ilic S., et al.* Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes // ACCV, 2012. Vol. 7724. Ch. 42. P. 548–562.

[30] *Harris C., Stephens M.* A combined corner and edge detector // Alvey Vision Conference. — Alvey Vision Club, 1988. P. 23.1–23.6.

[31] *Beyer H.* Advances in characterization and calibration of digital imaging systems // Int. Arch. Photogramm., 1992. Vol. XXIX. P. 545–555.

[32] *Lourakis M. I. A., Argyros A. A.* Sba: A software package for generic sparse bundle adjustment // ACM T. Math. Software, 2009. Vol. 36. No. 1. Article No. 2. doi: 10.1145/1486525.1486527.

[33] *Szeliski R.* Computer vision: Algorithms and applications. — 1st ed. — Texts in computer science ser. — London: Springer-Verlag, 2011. 832 p.

Поступила в редакцию 18.09.2017

Deep convolutional autoencoders: Stereo matching for three-dimensional model reconstruction of low-textured objects*

***V. V. Kniaz*[1,2], *O. V. Vygolov*[1], *V. V. Fedorenko*[1], *and V. D. Sevrykov*[1]**

vl.kniaz@gosniias.ru; o.vygolov@gosniias.ru; vfedorenko@gosniias.ru; vsevryukov@gosniias.ru

[1]FSUE State Scientific Research Institute of Aviation Systems, 7 Viktorenko Str., Moscow, Russia

[2]Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow, Russia

*The research was supported by the Russian Foundation for Basic Research (grants 17-29-04410 and 16-08-01260).

Methods: A new method for stereo matching based on deep neural convolutional autoencoders is presented. An autoencoder reduces the image dimensions and produces the code that could be used to perform an effective search of the corresponding image patch for low-textured object.
Results: An architecture of a new autoencoder was developed. The autoencoder performs coding and decoding of color images with resolution 32×32 pixels. A comparison of the performance of the developed method and modern image patch descriptors is presented. The method was applied to process images and to reconstruct three-dimensional (3D) models of archaeological excavations organized by the Bosphorus expedition of the Russian State Historical Museum.
Concluding Remarks: The analysis of an application of the developed method proves that it outperforms the existing image descriptors in the matching of image patches of low-textured objects.
Keywords: *deep learning; structure from motion; autoencoder*

DOI: 10.21469/22233792.3.2.03

References

[1] Lowe, D. G. 1999. Object recognition from local scale-invariant features. *Conference (International) on Computer Vision Proceedings*. Washington, DC: IEEE Computer Society. 1150. Available at: `http://dl.acm.org/citation.cfm?id=850924.851523` (accessed December 29, 2017).

[2] Bay, H., T. Tuytelaars, and L. Van Gool. 2006. Surf: Speeded up robust features. *ECCV*. 404–417.

[3] Klein, G., and D. W. Murray. 2007. Parallel tracking and mapping for small AR workspaces. *ISMAR*. 225–234.

[4] Engel, J., T. Schöps, and D. Cremers. 2014. LSD-SLAM: Large-scale direct monocular SLAM. *ECCV* 8690(Ch. 54):834–849.

[5] Kehl, W., F. Milletari, F. Tombari, S. Ilic, and N. Navab. 2016. Deep learning of local RGB-D patches for 3D object detection and 6D pose estimation. *ECCV* 9907(7):205–220.

[6] Rublee, E., V. Rabaud, K. Konolige, and G. Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. *Conference (International) on Computer Vision*. IEEE. 2564–2571.

[7] Remondino, F., M. G. Spera, E. Nocerino, F. Menna, and F. Nex. 2014. State of the art in high density image matching. *Photogramm. Rec.* 29(146):144–166.

[8] Knyaz, V. A., and A. G. Chibunichev. 2016. Photogrammetric techniques for road surface analysis. *Int. Arch. Photogramm.* XLI-B5:515–520. doi: 10.5194/isprs-archives-XLI-B5-515-2016. Available at: `https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLI-B5/515/2016/` (accessed December 29, 2017).

[9] Yamaguchi, M., H. Saito, and S. Yachida. 2017. Application of LSD-SLAM for visualization temperature in wide-area environment. *VISIGRAPP*. 216–223.

[10] Gallup, D., J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. 2007. Real-time plane-sweeping stereo with multiple sweeping directions. *CVPR*. 1–8.

[11] Bleyer, M., C. Rhemann, and C. Rother. 2011. PatchMatch stereo — stereo matching with slanted support windows. *British Machine Vision Conference*. British Machine Vision Association. 14.1–14.11.

[12] Galliani, S., K. Lasinger, and K. Schindler. 2016. Gipuma: Massively parallel multi-view stereo reconstruction. *Publ. Deutschen Gesellschaft Photogrammetrie Fernerkundung Geoinformation e. V* 25:361–369.

[13] Wohlhart, P., and V. Lepetit. 2015. Learning descriptors for object recognition and 3D pose estimation. arXiv.org. arXiv:1502.05908.

[14] Simo-Serra, E., E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. 2015. Discriminative learning of deep convolutional feature point descriptors. *ICCV*. 118–126.

[15] Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. 1097–1105.

[16] Szegedy, C., W. Liu, Y. Jia, *et al.* 2015. Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 1–9.

[17] Esteban, C. H., and F. Schmitt. 2003. Silhouette and stereo fusion for 3D object modeling. *4th Conference (International) on 3D Digital Imaging and Modelling Proceedings*. Banff, Alberta, Canada. 46–53.

[18] Vogiatzis, G., C. Hernandez, P. H. S. Torr, and R. Cipolla. 2007. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE T. Pattern Anal.* 29(12):2241–2246.

[19] Bethmann, F., and T. Luhmann. 2015. Semi-global matching in object space. *Int. Arch. Photogramm.* XL-3/W2:23–30.

[20] Hajebi, K., and J. S. Zelek. 2008. Structure from infrared stereo images. *Canadian Conference on Computer and Robot Vision Proceedings*. IEEE. 105–112.

[21] Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep learning*. The MIT Press. 800 p.

[22] Turchenko, V., and A. Luczak. 2015. Creation of a deep convolutional auto-encoder in caffe. *CoRR abs/1501.02565* 1512. arXiv:1512.01596.

[23] LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86(11):2278–2324.

[24] Krizhevsky, A., and G. Hinton. 2009. Learning multiple layers of features from tiny images. University of Toronto. Technical Report.

[25] Ridgeway, K., J. Snell, B. Roads, R. S. Zemel, and M. C. Mozer. 2015. Learning to generate images with perceptual similarity metrics. *CoRR*. arXiv:1511. 06409. Available at: `http://arxiv.org/abs/1511.06409` (accessed December 29, 2017).

[26] LeCun, Y., L. Bottou, G. B. Orr, and K. R. Müller. 1998. Efficient BackProp. *Neural networks: Tricks of the trade*. Eds. G. B. Orr and K.-R. Müller. Lecture notes in computer science ser. Berlin–Heidelberg: Springer. 1524:9–50.

[27] Zhang, R., P. Isola, and A. A. Efros. 2016. Colorful image colorization. *ECCV* 9907(Ch. 40):649–666.

[28] Knyaz, V. A. 2010. Multi-media projector — single camera photogrammetric system for fast 3D reconstruction. *Int. Arch. Photogramm.* XXXVIII-5:343–348. Available at: `http://www.isprs.org/proceedings/XXXVIII/part5/papers/143.pdf` (accessed December 29, 2017).

[29] Hinterstoisser, S., V. Lepetit, S. Ilic, *et al.* 2012. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. *ACCV* 7724(Ch. 42):548–562.

[30] Harris, C., and M. Stephens. 1988. A combined corner and edge detector. *Alvey Vision Conference*. Alvey Vision Club. 23.1–23.6.

[31] Beyer, H. 1992. Advances in characterization and calibration of digital imaging systems. *Int. Arch. Photogramm.* XXIX:545–555.

[32] Lourakis, M. I. A., and A. A. Argyros. 2009. Sba: A software package for generic sparse bundle adjustment. *ACM T. Math. Software* 36(1). Article No. 2. doi: 10.1145/1486525.1486527.

[33] Szeliski, R. 2010. *Computer vision: Algorithms and applications*. 1st ed. Texts in computer science ser. London: Springer-Verlag. 832 p.

Received September 18, 2017