СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ

УДК 004.93, 004.89

СИСТЕМА РАСПОЗНАВАНИЯ СТАТИЧЕСКИХ ЖЕСТОВ РУК С ИСПОЛЬЗОВАНИЕМ КАМЕРЫ ГЛУБИНЫ

Д. Ж. Сатыбалдина, к.ф.-м.н., профессор кафедры ВТ ЕНУ им. Л.Н.Гумилёва, Нур-Султан, Казахстан;
orcid.org/0000-0003-0291-4685, e-mail: satybaldina_dzh@enu.kz
Г. В. Овечкин, д.т.н., профессор, заведующий кафедрой ВПМ РГРТУ, Рязань, Россия;
orcid.org/0000-0001-6887-2217, e-mail: g_ovechkin@mail.ru
К. А. Калымова, докторант ЕНУ им.Л.Н.Гумилёва, Нур-Султан, Казахстан;
orcid.org/0000-0003-0610-740X, e-mail: gulzia_kalymova@mail.ru

Целью работы *является разработка системы для распознавания статических жестов рук на основе глубокой сверточной нейронной сети с использованием концепции трансфера обучения. Система распознавания жестов состоит из устройства захвата жестов (сенсора), алгоритмов предобработки и сегментации изображения, блока извлечения признаков и классификации жестов. Программная реализация выполнена с использованием инструментов Python 3.6. В качестве сенсора применяется камера глубины Intel® RealSense™ D435. Несколько Python-библиотек с открытым исходным кодом обеспечивают надежную реализацию алгоритмов обработки и сегментации изображений. Подсистема извлечения признаков и классификации жестов построена на основе архитектуры нейронной сети VGG-16, реализованной с использованием фреймворков глубокого обучения TensorFlow и Keras. Экспериментальные результаты показывают, что модель нейронной сети, обученная на базе данных жестов рук, состоящей из 2000 изображений, обеспечивает высокую точность распознавания на этапе тестирования системы.*

Ключевые слова: *интеллектуальные информационные системы и технологии, глубокое обучение, камера глубины, распознавание жестов, сверточная нейронная сеть, Keras, OpenCV, Python, TensorFlow, VGG-16.*

DOI: 10.21667/1995-4565-2020-72-93-105

Введение

В настоящее время трехмерные оптические системы стали популярны во многих областях применения: от робототехники [1] и автомобилестроения [2] до биомедицины [3]. Их успех обеспечивается в основном благодаря недавним разработкам, которые позволили создавать камеры глубины, являющиеся относительно недорогими, но точными и компактными. Видеопотоки с камер глубины подобны цветным, за исключением того, что каждый пиксель имеет значение, представляющее расстояние от камеры, а не информацию о цвете [4].

Впервые технология проецирования точечного рисунка на объект с использованием инфракрасного (ИК) проектора и обнаружения точек с помощью ИК-камеры для оценки глубины была применена израильской компанией PrimeSense [5]. Данный подход был реализован в сенсоре Kinect, первая версия которого была представлена в сентябре 2009 года. Результатом работы Kinect была карта глубины с разрешением 320×240.

После того, как в конце 2013 года компания Apple Inc. стала правообладателем компании PrimeSense, технология глубинных камер продолжила свое развитие. Во второй версии сен-

сора Kinect для XboxOne была использована ToF-технология (time-of-flight), которая отличается улучшенными параметрами по точности и разрешению [6]. Сенсор глубины, встроенный в смартфон iPhone X, используется для распознавания лиц в реализации функции Apple FaceID [6].

Компания Intel также создала собственную технологию глубинного зрения Intel RealSense и в сотрудничестве с корпорацией Microsoft разработала средство WindowsHello для 3D распознавания лиц, предоставляющее допуск в устройства Windows 10 [7]. Новые камеры глубины серии D400 были представлены в январе 2018 года. Они используют стереозрение для вычисления глубины [8]. Технология RealSense поддерживается комплектом разработки мультиплатформенного программного обеспечения с открытым исходным кодом [9].

Несмотря на то, что устройства Intel RealSense появились на рынке только в последние годы, они стали применяться во многих областях, например, в системах безопасности [10], робототехнике [11], медицине [12], сельском хозяйстве [13]. В то же время имеются единичные работы, в которых сообщается об использовании этих камер глубины для распознавания жестов рук, которые могут быть интегрированы в системы эффективного человеко-компьютерного взаимодействия.

Авторы одной из ранних работ [14] оценили влияние информации о глубине в процессе распознавания жестов и пришли к выводу, что использование силуэтов глубины значительно повышает точность распознавания. В работе [15] предложен для распознавания жестов рук алгоритм объединения нескольких глубинных дескрипторов. Результаты этих работ легли в основу ряда систем распознавания жестов рук с использованием камер глубины [16-21], характеристики которых представлены в таблице 1.

Как видно из таблицы 1, использованные в работах методы выделения признаков и классификации жестов обеспечивают высокую точность распознавания как статических, так и динамических жестов. Следует отметить, что в большинстве работ эффективность систем распознавания оценивается по небольшому набору жестов, что ограничивает их массовое применение. Это подтверждается тем фактом, что имеется множество разработок с открытым исходным кодом для распознавания лица, рта и глаз (OpenCV), но нет надежных детекторов рук. Поэтому остается актуальной задача совершенствования методов распознавания статических и динамических жестов рук на фотоизображении или в видео потоке.

Дополнительно отметим, что одной из тенденций последних лет является развитие и использование технологии глубокого обучения (*Deep Learning,* DL), которая используется в области цифровой обработки изображений для решения сложных задач (классификация, сегментация и обнаружение изображений). Методы DL, использующие сверточные нейронные сети (Convolutional Neural Network, CNN), уже оказали влияние на широкий спектр работ по обработке сигналов в рамках традиционных и новых областей, включая ключевые аспекты машинного обучения и искусственного интеллекта [22].

В связи с этим целью настоящей работы является создание новой системы для распознавания жестов, основанной на комбинированном использовании глубинного сенсора RealSense D435 для захвата жеста и предварительно обученной сверточной нейронной сети с архитектурой VGG-16 для выделения признаков и классификации объектов. Для программной реализации системы распознавания жестов на языке программирования Python использованы библиотеки RealSense от компании Intel, OpenCV и DL-фреймворки с открытым исходным кодом Keras и TensorFlow. Для тонкой настройки выходных слоев предобученной нейронной сети мы сформировали базу данных из 2000 изображений, состоящую из 40 различных видов изображений 5 жестов, которые показывали перед сенсором 10 человек. В целях тестирования и определения производительности предлагаемого подхода видеоданные от других участников эксперимента, не участвовавших в обучении нейронной сети, подавались на вход системы распознавания жестов напрямую с глубинной камеры.

Остальная часть этой статьи структурирована следующим образом. В разделе 2 представлена блок-схема алгоритма работы системы распознавания статических жестов рук и

кратко описаны основные этапы ее работы. В разделе 3 кратко показаны образцы жестов, база данных для обучения и тестирования приложения, выполнен анализ результатов эксперимента с точки зрения точности распознавания по сравнению с известными результатами. Заключение и направления будущих исследований представлены в разделе 4.

Таблица 1 – Сравнение методов распознавания жестов на основе использования камер глубины
Table 1 – Comparison of gesture recognition methods based on depth cameras application

Работа	Статический (S) или динамический (D) жест	Сенсор	Признаки распознавания	Метод классификации	Точность распознавания, %
[16]	D	CSEM Swissranger SR-2	Примитивы движений	Классификатор вероятностного редактирования расстояния	92,9
[17]	S	ToF и RGB камеры	Трехмерные модели положений рук	Метод ближайших соседей (Nearest Neighbors)	99,07
[18]	S	Kinect	Формы рук и пальцев	Сопоставление шаблонов с использованием FEMD (finger earth mover's distance)	93,9
[19]	D	Kinect	Изображение истории движений (Motion History Image)	Метод на основе вычисления максимума коэффициента корреляции	Не сообщается
[20]	S, D	Kinect	Значения пикселей глубины	*Классификация*, основанная на принципах *случайного леса*	91,84
[21]	S	RealSense SR300	Значения пикселей глубины и пикселей цвета	Сверточная нейронная сеть, содержащая два входных канала: цветные изображения и изображения глубины	99,4

Предлагаемый подход

Система распознавания статических жестов рук

Жест – это конфигурация и / или движение части тела, выражающая эмоцию, намерение или команду. Набор жестов и их значения образуют словарный запас жестов.

Жесты можно разделить на два типа: статические и динамические. В статических жестах положение руки не изменяется во время демонстрации жеста. Статические жесты в основном зависят от формы ладони и углов поворота пальцев рук и целой кисти. Во втором случае положение руки меняется непрерывно и динамические жесты зависят от траекторий и ориентаций рук в дополнение к форме и углам поворота пальцев [23].

Распознавание жестов рук является сложной задачей из-за как объективных, так и субъективных различий, связанных с большим количеством степеней свободы костей кисти и пальцев, различиями в артикуляции, относительно малой площадью области рук, разным цветом кожного покрова. Кроме того, надежные алгоритмы сегментации и детектирования положений рук и пальцев должны иметь инвариантность относительно размера, скорости и ориентации жеста, освещенности сцены, неоднородности фона и других параметров. Таким образом, автоматизация распознавания жестов рук связана с решением ряда сложных задач.

Для решения вышеупомянутых задач в настоящей работе предлагается система распознавания статических жестов рук, основанная на цифровой обработке цветных и глубинных изображений кадров из видеопотока в режиме реального времени и извлечении из них клас-

сифицирующих признаков внутри сверточной нейронной сети, предобученной на базе данных изображений ImageNet. Блок-схема алгоритма работы предлагаемой системы распознавания статических жестов рук изображена на рисунке 1. Каждый этап работы данной системы кратко описан в следующих подразделах.

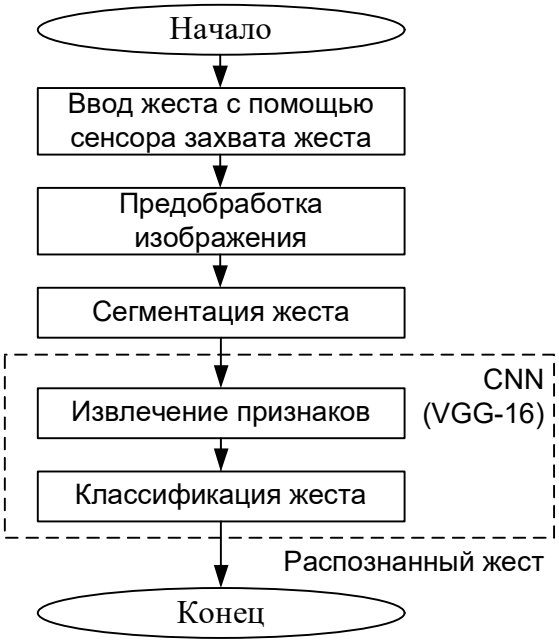

Рисунок 1 – Блок-схема алгоритма работы предлагаемой системы распознавания статических жестов рук
Figure 1 – Block diagram of static hand gesture recognition system proposed

Захват изображений с жестами рук

Назначение сенсора захвата жестов – преобразовать жест в цифровую форму. В качестве устройства сбора данных используется глубинная камера Intel® RealSense ™ D435, которая дает RGB-изображение, а также глубину для каждой точки [8].

RealSense D435 – это компактное (99 mm × 25 mm × 25 mm; вес 72 г.) периферийное RGB-D устройство с поддержкой стандарта USB 3.1 и радиусом действия до 10 м. Камера состоит из процессора визуализации D4, глубинного модуля (Depth Module) и RGB-камеры. Характеристики глубинного и RGB разрешения – до 1280×720 и 1920×1080 соответственно. Для реализации стереозрения использовались два одинаковых сканера (левый и правый) и инфракрасный проектор (рисунок 2). Инфракрасный проектор проецирует невидимый статический шаблон на сцену с низкой текстурой. Левый и правый сканеры снимают сцену и отправляют данные визуализации в процессор глубинной визуализации, который вычисляет значения глубины для каждого пикселя в изображении, сопоставляя точки на левом изображении с правым изображением [8]. Значения пикселей глубины обрабатываются для создания кадра глубины. Последовательные кадры глубины создают видеопоток глубины.

В открытой библиотеке RealSense SDK 2.0 имеются стандартные функции для инициализации камеры, установки параметров ее работы, функции и методы чтения кадров из видеопотока, вычисления расстояния от руки до глубинной камеры, методы сохранения RGB-изображений и карт глубины [9]. Есть возможность модифицировать алгоритмы, доступные в исходных кодах RealSense SDK 2.0. Для реализации захвата кадров с жестами использованы методы и функции из библиотеки RealSense SDK 2.0, а также созданные авторами дополнительные функции для захвата жестов.

а (a) **б(b)** **в(c)**

Рисунок 2 – Общий вид камеры RealSense D435 (а) и примеры данных с сенсора: RGB - изображение (б) и кадр глубины (в)
Figure 2 –RealSense D435 (a) and output data from the sensor: RGB Image (b) and Depth Image (c)

Предварительная обработка изображений

Выполнение различных операций на последующих этапах работы системы распознавания жестов, таких как сегментация и извлечение признаков, намного проще для предварительно обработанных изображений. На данном этапе обычно помехи и внешние шумы уменьшаются применением операций усреднения и выравнивания гистограмм, проводится также цветовая нормализация в соответствии с условиями освещения и световой температурой [15].

В данной работе для удаления шумов в кадре применяется двусторонний фильтр (bilateral filter) из библитеки OpenCV. Авторами также реализован дополнительный метод определения средней яркости пикселей в кадре [24].

Методы сегментации

Сегментация – это извлечение интересующего объекта (руки) из фона и определение его местоположения в сцене. На данном шаге выполняются следующие операции (рисунок 3):

– поиск области интереса (Region of Interest, ROI) на изображении для обнаружения руки и удаления фона;

– преобразование цветного изображения в оттенки серого;

– применение фильтра Гаусса для удаления шумов в двумерных ихображениях в оттенках серого;

– построение контуров сегментированного объекта;

– пороговое преобразование для получения сегментированного жеста руки.

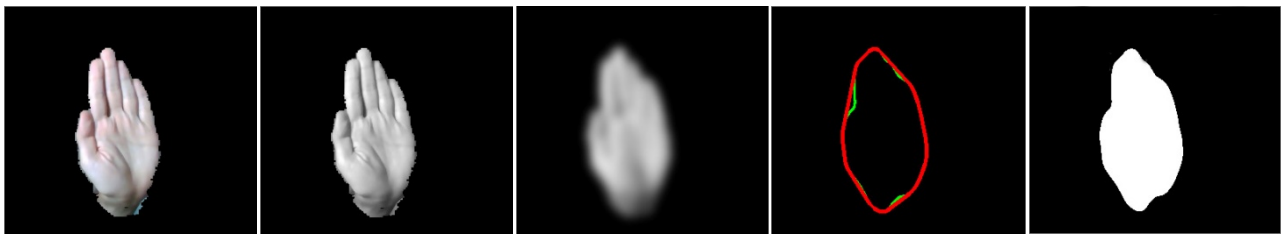

Рисунок 3 – Сегментирование жеста рук
Figure 3 – Hand Gesture Segmentation

Для реализации этих операций использованы как методы из библитеки OpenCV, так и методы, реализованные авторами самостоятельно.

Для примера рассмотрим более подробно метод, использованный для детектирования руки в кадре и удаления фона. При сравнении кадров используются данные с камеры глубины. Изображение глубины проецируется на изображение RGB, что дает точное значение глубины для каждого пикселя в изображении RGB. Вычитание заднего фона основано на сравнении пикселей глубины с порогом clipping_distance: если значение больше порога, то в результирующем кадре пиксель закрашивается черным цветом, в противном случае – цветом текущего RGB-кадра. Функции обнаружения руки основаны на вычислении необходимого

количество пикселей в кадре (значение определяется экспериментально) без заднего фона в зафиксированной зоне. Так как на кадре весь задний фон закрашен черным, то можно просто просчитывать количество пикселей, которые не равны нулю. Если набирается необходимое количество пикселей, то система фиксирует руку в кадре.

Базовая архитектура VGG-16 и трансфер обучения

Механизм классификации или распознавания – это вычислительный алгоритм, который принимает представление объекта и соотносит (классифицирует) его как экземпляр некоторого известного типа класса. В предлагаемом подходе по распознаванию статических жестов использована глубокая сверточная нейронная сеть (Deep Convolutional Neural Network, DCNN) с архитектурой VGG-16, предварительно обученная на большом наборе изображений.

Архитектура VGG-16 на основе глубокого обучения была развита авторами работы [25]. Созданная ими последовательная структура архитектуры с небольшими фильтрами свертки размером 3 × 3 показала, что можно значительно улучшить работу нейронных сетей, увеличив глубину до 16-19 взвешенных слоев. Базовая модель VGG-16 показана на рисунке 4, *а*. За 5-ю сверточными блоками следуют 5 слоев подвыборки и 3 полносвязанных слоя. Для выходного слоя с 1000 выходными каналами используется функция активации *softmax*, которая вычисляет распределение вероятности для 1000 классов объектов. МодельVGG-16 была обучена и протестирована на базе данных ImageNet для классификации 1,3 миллиона изображений на 1000 классов, достигнутая точность составила 92,7 % [25].

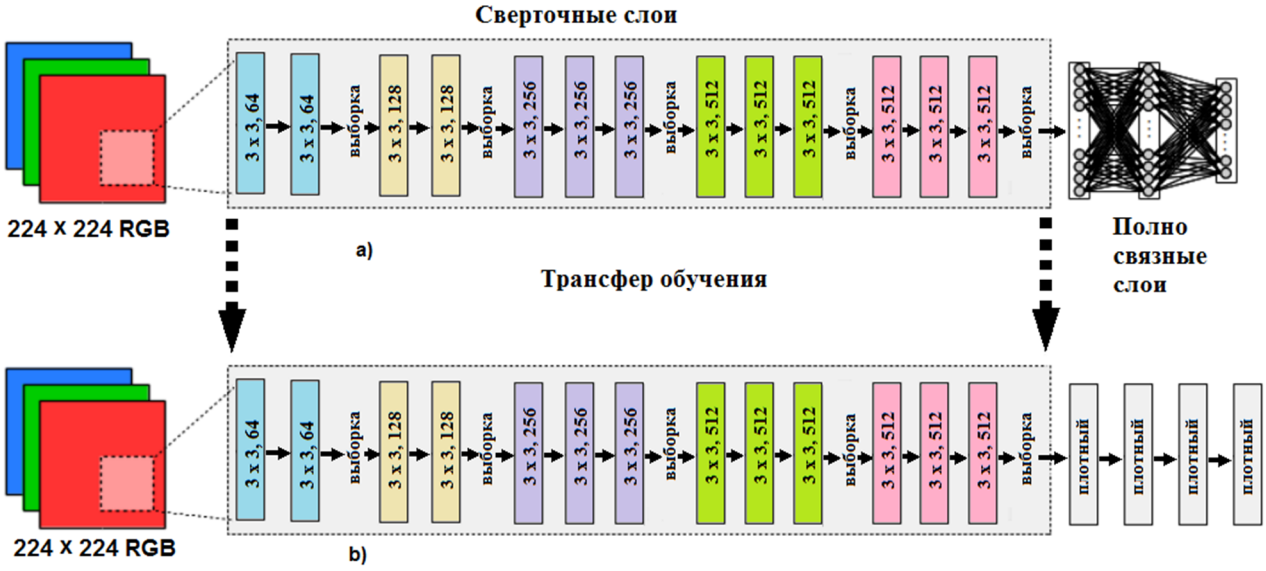

Рисунок 4 – Архитектура VGG16 с весами, предварительно обученными в ImageNet (а) и предлагаемая модель DCNN (б)
Figure 4 – VGG16 architecture with weights pre-trained on ImageNet (a) and offered DCNN model (b)

Использование DCNN с фиксированными весами сверточных слоев, полученных при обучении на крупномасштабных наборах естественных изображений (ImageNet), показывает высокую точность классификации изображений из других предметных областей (например, байт-плотов вредоносных программ) с помощью концепции трансфера обучения (a transfer learning) и «тонкой настройки» (a fine-tuning) [26]. Трансфер обучения заключается в переносе параметров нейронной сети, обученной с одним набором данных и задачей, в другую задачу классификации с другим набором изображений [27]. «Тонкая настройка» связана с необходимостью изменения и переобучения полносвязанных слоев в классификаторе. Базовая архитектура VGG-16 содержит 1000 выходных нейронов в последнем слое по количеству классов объектов. В новой задаче классификации (например, в задаче распознавания жестов) количество классов может отличаться от того, которое было в исходном наборе данных. В

таком случае последний слой в архитектуре VGG-16 должен быть удален, а добавлен новый классификатор с необходимым количеством выходных нейронов. Необходимо также заменить предыдущие полносвязанные слои, так как и их выходные векторы не соответствуют новому классифицирующему слою.

Весовые коэффициенты новых слоев инициализируются случайными значениями. После этого начинается процесс их обучения на обучающем наборе данных. Мы применили стратегию обучения из конца в конец (end-to-end), при которой предобученные весовые коэффициенты не фиксируются, а корректируются под обучающий набор данных, т.е. поддаются «тонкой настройке».

В предлагаемой системе распознавания статических жестов рук используется модифицированная архитектура VGG-16, в которой полносвязанные слои с большим количеством нейронов заменены на 4 плотных слоя с меньшим количеством нейронов (рисунок 4, *б*). В последнем слое имеется 5 выходных каналов для 5 жестов из обучающего и тестового наборов жестов. Входные данные DCNN с архитектурой VGG-16 представляют собой RGB изображение фиксированного размера 224 × 224. Поэтому на этапе сегментации обработанные кадры с сегментированным жестом, сохраненные в цветовой модели RGB с другим разрешением, преобразуются в формат 224 × 224 и передаются на вход модифицированной модели VGG-16. Пройдя через несколько слоев свертки и выборки, RGB изображения жестов рук преобразуются в карты объектов и отправляются в плотные слои. На последнем шаге используется модель *softmax* для прогнозирования жеста и вывода результата в качестве вероятности классификации.

Экспериментальные исследования

База данных статических жестов рук

Мы подготовили базу данных, которая содержит изображения с сегментированными жестами, представленными на рисунке 5. Эти жесты включены в альтернативный набор данных [28], который можно использовать для сравнения или перекрестной проверки предложенной системы распознавания статических жестов рук.

Камеру глубины Intel RealSense D435 разместили на столе, участники эксперимента представляли жесты в положении сидя, располагая руку перед сенсором на расстоянии от 10 до 75 см перед ним (рисунок 6). Жесты выполнялись 10-ю различными субъектами.

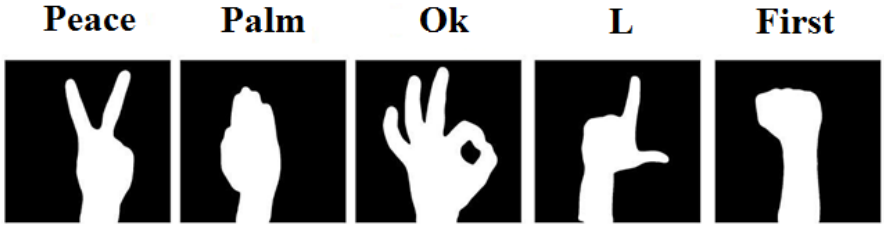

Рисунок 5 – Образцы жестов из базы данных
Figure 5 – Dataset samples

Наш набор данных имеет в общей сложности 2000 изображений, в том числе 1000 изображений RGB и 1000 карт глубины, собранных при разных фонах в нескольких комнатах с изменениями освещенности. Чтобы увеличить разнообразие базы данных, мы также пытались увеличить количество изображений с помощью вариации расстояния до сенсора, углов наклона, поворота ладони и т.д. Собранная база статических жестов рук использовалась только для обучения модифицированной сети VGG-16. На этапе тестирования предлагаемой системы изображения RGB и пиксели глубины с сенсора RealSense D435 используются в режиме реального времени для классификации жестов.

Обучение глубокой нейронной сети – это процесс «тонкой настройки», основанный на моделях VGG-16, предварительно обученных на базе изображений ImageNet. В начале скорость обучения составляет 0,01 и уменьшается в 10 раз каждые 2000 итераций. Уменьшение

веса установлено на 0,0004. Для «тонкой настройки» тренировочного набора потребовалось не более 10 000 итераций. Эксперименты проводились на процессоре Intel® Core(TM) i7-9750H CPU, NVIDIA GeForce GTX 1650, 16 GB RAM.

Рисунок 6 – Пример захвата жеста руки человека сенсором RealSense D400
Figure 6 – Sample of human hand gesture captured by RealSense D400 sensor

Показатели эффективности

Для получения количественных показателей эффективности предложенного подхода используется матрица ошибок (a confusion matrix, CM). Показатели точности (a precision), полноты (a recall) и F-мера также применяются в качестве метрик для оценки производительности подхода.

Матрица CM размером $n \times n$ содержит информацию о том, сколько раз система классификации приняла правильное и сколько раз неверное решение по объектам заданного класса, где n – количество различных классов [29]. В таблице 2 приведен пример матрицы ошибок для задачи классификации на два класса.

Таблица 2 – Матрица ошибок для задачи бинарной классификации
Table 2 – Confusion matrix for two-class classification problem

	Классифицированный	
Фактический	Negative	Positive
Negative	*a*	*c*
Positive	*d*	*b*

В ячейках таблицы находятся следующие значения:
– *a*, количество истинно-отрицательных решений;
– *b*, количество истинно-положительных решений;
– *c*, количество ложно-отрицательных решений;
– *d*, количество ложно-положительных решений.

В работе мы используем альтернативную конструкцию CM, в которой каждый столбец представляет процент жестов, принадлежащих каждому классу, и вдоль первой диагонали указаны максимальные значения вероятности правильной идентификации жеста, т.е. отнесения наблюдаемого объекта к некоторому классу жестов. Недиагональные элементы матрицы могут содержать значение вероятности классификации, сравнимое с некоторым пороговым значением. Если вычисленная вероятность классификации больше этого порога или равна ему, то предсказание считается правильным, в противном случае – неверным [29]. Данный вариант матрицы ошибок соответствует формату выходных данных со сверточной нейронной сети, в финальном слое которой используется функция softmax, вычисляющая результат прогнозирования жеста в виде вероятности классификации (таблица 3).

Таблица 3 – Матрица ошибок для задачи распознавания статических жестов рук
Table 3 – Confusion matrix for the proposed static hand gesture recognition system

	Решение классификатора					
Фактический жест	First	L	Ok	Palm	Peace	$Precision_c$
First	99,9798	0,0018	0,0131	0,005	0,0003	0,9998
L	0	99,9978	0,0021	0	0,0001	1,0000
Ok	0,0008	0,0255	99,9726	0	0,0011	0,9997
Palm	0,0194	0,0159	0,0151	97,4405	2,5091	0,9744
Peace	0	0,0004	0,0032	0,6245	99,3719	0,9937
$Recall_c$	0,9998	0,9996	0,9997	0,9936	0,9754	

Из СМ-матрицы можно непосредственно вычислить показатели точности и полноты, которые можно выразить следующим образом:

$$\boldsymbol{Precision}_{\mathrm{c}} = \frac{\boldsymbol{A_{c,c}}}{\sum_{i=1}^{n} \boldsymbol{A_{c,i}}}, \quad (1)$$

$$\boldsymbol{Recall}_{\mathrm{c}} = \frac{\boldsymbol{A_{c,c}}}{\sum_{i=1}^{n} \boldsymbol{A_{i,c}}}. \quad (2)$$

Из формул (1), (2) видно, что точность определяется отношением соответствующего диагонального элемента матрицы к сумме всей строки данного класса, а полнота – отношением диагонального элемента матрицы к сумме всего столбца класса.

Показатель полноты (*Recall*) демонстрирует способность алгоритма обнаруживать данный класс в целом, а точность (*Precision*) демонстрирует способность отличать этот класс от других классов. Показатель *F-Score* – гармоническое среднее точности и полноты:

$$\boldsymbol{F\text{-}Score} = 2 \times \frac{\boldsymbol{Precision_c} \times \boldsymbol{Recall_c}}{\boldsymbol{Precision_c} + \boldsymbol{Recall_c}}. \quad (3)$$

Метрика *F-Score* достигает своего максимума с показателями точности и полноты, равными единице, и близка к нулю, если один из аргументов близок к нулю.

Результаты и обсуждение

В таблице 3 представлена матрица средних вероятностей классификаций, полученных с использованием предложенной системы распознавания статических жестов на стадии тестирования. В эксперименте по тестированию системы участвовало 5 человек, каждый из которых перед камерой глубины продемонстрировал по 40 различных представлений 5-и классов жестов. Таким образом, общее количество образцов жестов на этапе тестирования составило 1000, что равно половине обучающей базы. Как видно, для всех жестов значение элементов главной диагонали составляет более 97 %, и лишь небольшой процент выборок определяется как принадлежащий другим жестам (менее 0,2 %). Это указывает на то, что предложенный подход имеет высокую производительность как по точности, так по полноте. Это подтверждается вычисленными значениями $Precision_c$ и $Recall_c$ (см. последний столбец и последнюю строку в таблице 3). Исключением является распознавание жеста «Palm», где ошибка классификации превышает 2,5 %, что можно объяснить определенным сходством этого жеста с жестом «Peace». Учитывая, что результирующая точность классификатора рассчитывается как среднее арифметическое его точности по всем классам, производительность предлагаемого подхода по точности распознавания получена на уровне 0,9935, или ≈99,4 %.

Эффективность предлагаемой системы распознавания жестов рук была рассмотрена в сравнении с системой, развитой авторами работы [26], в которой для захвата изображений с жестами был использован сенсор Leap Motion. Для сравнения результатов обоих решений используется показатель *F-Score* (см. Таблицу 4).

Как видно из таблицы 4, наша система обеспечивает получение сопоставимых результатов. Подход распознавания жестов рук, предложенный в [26], использует RGB-сенсор и глобальный дескриптор изображения на основе глубинных пространственных диаграмм квантованных паттернов (Depth Spatiograms of Quantized Patterns, DSQP) без какой-либо стадии

сегментации руки. Дескрипторы изображений, уменьшенные из DSQP, анализируются набором машин векторов поддержки для классификации жестов. Это сравнение также позволяет нам заметить, что карты глубины и сверточные нейронные сети позволяют достичь высокой производительности без дополнительных процедур обработки.

Таблица 4 – Сравнение показателей *F-Score* для двух систем распознавания жестов
Table 4 – Comparison of *F-Score* metrics for two gesture recognition systems

	Система распознавания жестов	
Жест	Предлагаемый подход	[28]
First	1	0,99
L	1	0,99
Ok	1	0,99
Palm	0,98	1
Peace	0,98	0,99

Заключение

В работе представлена система распознавания статических жестов рук, которая использует пиксели цвета и глубины от сенсора Intel® Real Sense ™ D435. Уникальные возможности камеры глубины использованы, чтобы сегментировать жесты рук и удалить фон на изображениях с жестом. Выделение признаков и классификация жестов были выполнены с использованием глубоких сверточных нейронных сетей архитектуры VGG, предварительно обученных на базе данных ImageNet. Предложенный подход реализован на языке Python с трансфером обучения нейронной сети с использованием фреймворков Keras и TensorFlow. Обучающая база данных была собрана вручную с использованием RGB-изображений и карт глубины от камеры глубины. Эта база данных состоит из 2000 образцов, собранных с привлечением 10 участников эксперимента. Модифицированная модель VGG-16 обеспечивает высокую точность на тренировочном наборе изображений. Полученные оценки распознавания на этапе тестирования для 1000 образцов жестов подтверждают эффективность предварительно представленной структуры распознавания и указывают на возможности камеры глубины D435 для будущих приложений на основе человеко-компьютерного взаимодействия.

Однако распознавания статических жестов рук недостаточно для эффективных систем взаимодействия людей и компьютерных систем, роботов. Поэтому будущие исследования связаны с разработкой методов распознавания динамических жестов рук, а также с расширением базы данных и классов объектов.

Работа выполнена при поддержке Министерства цифрового развития, инноваций и аэрокосмической промышленности Республики Казахстан в рамках проекта № AP06850817.

Библиографический список

1. **Sturm J., Engelhard N., Endres F., Burgard W., Cremers D.** A benchmark for the evaluation of RGB-D SLAM systems // In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura, Portugal, 7-12 October 2012. Pp. 573-580.

2. **Makris S., Tsarouchi P., Surdilovic D., Krüger J**. Intuitive dual arm robot programming for assembly operations // CIRP Ann. 2014, vol. 63, pp. 13-16.

3. **Chiu C. Y., Thelwell M., Senior T., Choppin S., Hart J., & Wheat J.** Comparison of Depth Cameras for 3D Reconstruction in Medicine // Journal of Engineering in Medicine. 2019, vol. 233 (9), pp. 938-947.

4. **Carfagni M., Furferi R., Governi L., Santarelli C., Servi M., Uccheddu F., & Volpe Y**. Metrological and critical characterization of the Intel D415 stereo depth camera // Sensors. 2019, vol. 19, no. 3, pp. 489-508.

5. PrimeSense. Available online: http://xtionprolive.com/ (accessed on 24 January 2020).

6. **Zhang S**. High-speed 3D shape measurement with structured light methods: A review // Opt. Lasers Eng. 2018, vol. 106, pp. 119-131.

7. **Keselman L., Woodfill J.I., Grunnet-Jepsen A., Bhowmik A**. Intel® RealSense™ Stereoscopic Depth Cameras // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017, pp. 1-10.

8. Intel RealSense D400 Series Product Family. Datasheet. 2019 Intel Corporation. Document Number: 337029-007. Available online: https://www.intel.com/ (accessed on 24 January 2020).

9. Intel® RealSense™ SDK 2.0. Available online: https://www.intelrealsense.com/developers/ (accessed on 24 January 2020).

10. **Bock R. D.** Low-cost 3D security camera // Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything. International Society for Optics and Photonics. 2018, vol. 10643, pp. 106430E.

11. **Fang Q., Kyrarini M., Ristic-Durrant D., Gräser A.** RGB-D Camera based 3D Human Mouth Detection and Tracking Towards Robotic Feeding Assistance // Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference. 2018, pp. 391-396.

12. **Aoki H., Suzuki A., Shiga T.** Study on Non-Contact Heart Beat Measurement Method by Using Depth Sensor // World Congress on Medical Physics and Biomedical Engineering 2018. Springer, Singapore, 2019, pp. 341-345.

13. **Syed T. N., Jizhan L., Xin Z., Shengyi Z., Yan Y., Mohamed S. H. A., Lakhiar I. A**. Seedling-lump integrated non-destructive monitoring for automatic transplanting with Intel RealSense depth camera // Artificial Intelligence in Agriculture. 2019, vol. 3, pp. 18-32.

14. **Munoz-Salinas R., Medina-Carnicer R., Madrid-Cuevas F.J., Carmona-Poyato A.** Depth silhouettes for gesture recognition // Pattern Recognit. Lett. 2008, vol. 29 (3), pp. 319-329.

15. **Dominio F., Donadeo M., Zanuttig P.** Combining multiple depth-based descriptors for hand gesture recognition // Pattern Recognit. Lett. 2014, vol. 50, pp. 101-111.

16. **Holte M. B., Moeslund T. B., Fihl P.** Fusion of range and intensity information for view invariant gesture recognition // 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2008, pp. 1-7.

17. **Van den Bergh M**. **et al**. Real-time 3D hand gesture interaction with a robot for understanding directions from humans // 2011 Ro-Man. IEEE. 2011, pp. 357-362.

18. **Ren Z., Yuan J., Zhang Z.** Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera // Proceedings of the 19th ACM international conference on Multimedia. 2011, pp. 1093-1096.

19. **Wu D., Zhu F., Shao L.** One shot learning gesture recognition from rgbd images // 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2012, pp. 7-12.

20. **Keskin C., Kirac F., Kara Y., Akarun L.** Randomized decision forests for static and dynamic hand shape classification // 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2012, pp. 31-36.

21. **Liao B., Li J., Ju Z., Ouyang G.** Hand gesture recognition with generalized hough transform and DC-CNN using realsense // 2018 Eighth International Conference on Information Science and Technology (ICIST). 2018, pp. 84-90.

22. **Sree S. R., Vyshnavi S. B., Jayapandian N.** Real-World Application of Machine Learning and Deep Learning // 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE. 2019, pp. 1069-1073.

23. **Pisharady P. K., Saerbeck M.** Recent methods and databases in vision-based hand gesture recognition: A review // Computer Vision and Image Understanding. 2015, vol. 141, pp. 152-165.

24. **Chernov V., Alander J., Bochko V.** Integer-based accurate conversion between RGB and HSV color spaces // Computers & Electrical Engineering. 2015, vol. 46, pp. 328-337.

25. **Simonyan K. and Zisserman A.** Very deep convolutional networks for large-scale image recognition // arXiv preprint arXiv:1409.1556. 2014.

26. **Rezende E., Ruppert G., Carvalho T., Theophilo A., Ramos F., & de Geus P.** Malicious software classification using VGG16 deep neural network's bottleneck features // Information Technology-New Generations. Springer, Cham. 2018, pp. 51-59.

27. **Liu Z., Wu J., Fu L., Majeed Y., Feng Y., Li R., & Cui Y**. Improved kiwifruit detection using pre-trained VGG16 with RGB and NIR information fusion // IEEE Access. 2019, pp.2327-2336.

28. **Mantecón T., del Blanco C.R., Jaureguizar F., García N.** Hand Gesture Recognition using Infrared Imagery Provided by Leap Motion Controller // International Conference on Advanced Concepts for Intelligent Vision Systems. Springer, Cham. 2016. LNCS 10016, pp. 47-57.

29. **Visa S., Ramsay B., Ralescu A. L., & Van Der Knaap E**. Confusion Matrix-based Feature Selection // MAICS. 2011, vol. 710, pp. 120-127.

UDC 004.93:004.89

STATIC HAND GESTURES RECOGNITION SYSTEM WITH USING DEPTH CAMERA

D. Zh. Satybaldina, Ph.D. (Phys. and Math.), professor of Computer Engineering Department, L.N.Gumilyov ENU, Nur-Sultan, Kazakhstan;
orcid.org/0000-0003-0291-4685, e-mail: satybaldina_dzh@enu.kz
G. V. Ovechkin, Dr. Sc. (Tech.), full professor, Head of the Department, RSREU, Ryazan, Russia;
orcid.org/0000-0001-6887-2217, e-mail: g_ovechkin@mail.ru
G. A. Kalymova, Ph.D. student, L.N.Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan;
orcid.org/0000-0003-0610-740X, e-mail: gulzia_kalymova@mail.ru

The aim of the work *is to develop a system for static hand gestures recognition based on a convolutional neural network using transfer learning framework. The gesture recognition system consists of a gesture capture device (sensor), preprocessing and image segmentation algorithms, a feature extraction and gestures classification block. This work is performed in Python 3.6 tools. As a sensor, Intel® RealSense™ depth camera D435 is used. Several Python libraries, which provide solid implementations of image processing and segmentation, are used. The subsystem for features extracting and gestures classification is based on the modified VGG-16, being realized with the help of TensorFlow & Keras deep learning frameworks. Experimental results show that the proposed model, trained on the database of 2000 images, provides high recognition accuracy on testing stage.*

Key words: *intelligent information systems and technologies, deep learning, depth camera, gesture recognition, convolutional neural network, Keras, OpenCV, Python, TensorFlow, VGG-16.*

DOI: 10.21667/1995-4565-2020-72-93-105

References

1. **Sturm J., Engelhard N., Endres F., Burgard W., Cremers D.** A benchmark for the evaluation of RGB-D SLAM systems. *In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vilamoura, Portugal, 7-12 October 2012. Pp. 573-580.

2. **Makris S., Tsarouchi P., Surdilovic D., Krüger J**. Intuitive dual arm robot programming for assembly operations. *CIRP Ann*. 2014, vol. 63, pp. 13-16.

3. **Chiu C. Y., Thelwell M., Senior T., Choppin S., Hart J., & Wheat J.** Comparison of Depth Cameras for 3D Reconstruction in Medicine. *Journal of Engineering in Medicine*. 2019, vol. 233 (9), pp. 938-947.

4. **Carfagni M., Furferi R., Governi L., Santarelli C., Servi M., Uccheddu F., & Volpe Y**. Metrological and critical characterization of the Intel D415 stereo depth camera. *Sensors*. 2019, vol. 19, no. 3, pp. 489-508.

5. PrimeSense. Available online: http://xtionprolive.com/ (accessed on 24 January 2020).

6. **Zhang S**. High-speed 3D shape measurement with structured light methods: A review. *Opt. Lasers Eng*. 2018, vol. 106, pp. 119-131.

7. **Keselman L., Woodfill J.I., Grunnet-Jepsen A., Bhowmik A**. Intel® RealSense™ Stereoscopic Depth Cameras. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 1-10.

8. Intel RealSense D400 Series Product Family. Datasheet. 2019 Intel Corporation. Document Number: 337029-007. Available online: https://www.intel.com/ (accessed on 24 January 2020).

9. Intel® RealSense™ SDK 2.0. Available online: https://www.intelrealsense.com/developers/ (accessed on 24 January 2020).

10. **Bock R. D.** Low-cost 3D security camera. *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything. International Society for Optics and Photonics*. 2018, vol. 10643, pp. 106430E.

11. **Fang Q., Kyrarini M., Ristic-Durrant D., Gräser A.** RGB-D Camera based 3D Human Mouth Detection and Tracking Towards Robotic Feeding Assistance // *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*. 2018, pp. 391-396.

12. **Aoki H., Suzuki A., Shiga T.** Study on Non-Contact Heart Beat Measurement Method by Using Depth Sensor // *World Congress on Medical Physics and Biomedical Engineering 2018*. Springer, Singapore, 2019, pp. 341-345.

13. **Syed T. N., Jizhan L., Xin Z., Shengyi Z., Yan Y., Mohamed S. H. A., Lakhiar I. A**. Seedling-lump integrated non-destructive monitoring for automatic transplanting with Intel RealSense depth camera. *Artificial Intelligence in Agriculture*. 2019, vol. 3, pp. 18-32.

14. **Munoz-Salinas R., Medina-Carnicer R., Madrid-Cuevas F.J., Carmona-Poyato A.** Depth silhouettes for gesture recognition. *Pattern Recognit. Lett*. 2008, vol. 29 (3), pp. 319-329.

15. **Dominio F., Donadeo M., Zanuttig P.** Combining multiple depth-based descriptors for hand gesture recognition. *Pattern Recognit*. Lett. 2014, vol. 50, pp. 101-111.

16. **Holte M. B., Moeslund T. B., Fihl P.** Fusion of range and intensity information for view invariant gesture recognition. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 2008, pp. 1-7.

17. **Van den Bergh M**. **et al**. Real-time 3D hand gesture interaction with a robot for understanding directions from humans. *2011 Ro-Man. IEEE*. 2011, pp. 357-362.

18. **Ren Z., Yuan J., Zhang Z.** Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. *Proceedings of the 19th ACM international conference on Multimedia*. 2011, pp. 1093-1096.

19. **Wu D., Zhu F., Shao L.** One shot learning gesture recognition from rgbd images. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 2012, pp. 7-12.

20. **Keskin C., Kirac F., Kara Y., Akarun L.** Randomized decision forests for static and dynamic hand shape classification. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 2012, pp. 31-36.

21. **Liao B., Li J., Ju Z., Ouyang G.** Hand gesture recognition with generalized hough transform and DC-CNN using realsense. *2018 Eighth International Conference on Information Science and Technology (ICIST)*. 2018, pp. 84-90.

22. **Sree S. R., Vyshnavi S. B., Jayapandian N.** Real-World Application of Machine Learning and Deep Learning. *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE*. 2019, pp. 1069-1073.

23. **Pisharady P. K., Saerbeck M.** Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding*. 2015, vol. 141, pp. 152-165.

24. **Chernov V., Alander J., Bochko V.** Integer-based accurate conversion between RGB and HSV color spaces. *Computers & Electrical Engineering*. 2015, vol. 46, pp. 328-337.

25. **Simonyan K. and Zisserman A.** Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556. 2014.

26. **Rezende E., Ruppert G., Carvalho T., Theophilo A., Ramos F., & de Geus P.** Malicious software classification using VGG16 deep neural network's bottleneck features. *Information Technology-New Generations*. Springer, Cham. 2018, pp. 51-59.

27. **Liu Z., Wu J., Fu L., Majeed Y., Feng Y., Li R., & Cui Y**. Improved kiwifruit detection using pre-trained VGG16 with RGB and NIR information fusion. *IEEE Access*. 2019, pp.2327-2336.

28. **Mantecón T., del Blanco C.R., Jaureguizar F., García N.** Hand Gesture Recognition using Infrared Imagery Provided by Leap Motion Controller. *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, Cham. 2016. LNCS 10016, pp. 47-57.

29. **Visa S., Ramsay B., Ralescu A. L., & Van Der Knaap E**. Confusion Matrix-based Feature Selection. *MAICS*. 2011, vol. 710, pp. 120-127.