ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ РАСПОЗНАВАНИЯ ЖЕСТОВ ДЛЯ ЧЕЛОВЕКО-МАШИННОГО ВЗАИМОДЕЙСТВИЯ НА БАЗЕ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ

Поткин О.А., Филиппович А.Ю.
Московский Политехнический Университет
Москва, Российская Федерация
olpotkin@gmail.com

Аннотация. В статье описывается программный комплекс для обнаружения, отслеживания и классификации статических жестов в видеопотоке с использованием методов компьютерного зрения и глубокого обучения. Набор данных, используемый для решения этой задачи, является оригинальным, состоит более чем из 2000 уникальных изображений и 10 классов. Программный комплекс включает в себя модуль обнаружения кистей рук по цветовой маске, модуль отслеживания жестов, модуль классификации статических жестов на базе сверточной нейронной сети, а также вспомогательный модуль предварительной обработки изображений и модуль расширения набора данных.

Ключевые слова: классификация, жесты, сверточные нейронные сети, компьютерное зрение, детектирование

Вступление

Роботизированные системы становятся незаменимыми в различных отраслях промышленности. В последнее время концепция взаимодействия "человек-робот" привлекла внимание исследователей. Множество примеров демонстрирует, что человек обладает несравненными навыками решения проблем в основном благодаря развитым сенсорно-моторным способностям, но имеет ограниченную силу и уровень точности [1]. Однако роботизированные системы обладают устойчивостью к усталости, высокой скоростью, точностью и производительностью, но при этом имеют существенные ограничения в гибкости. Четко определенная и реализованная концепция взаимодействия "человек-робот" может освободить человека от тяжелых задач благодаря интуитивному и надежному интерфейсу.

Жесты являются одним из способов обмена информацией, общения. Информация, передаваемая мимикой и жестами, лежит в основе эффективного канала человеческой коммуникации [2]. Чтобы взаимодействовать с людьми,

роботизированные системы должны правильно понимать человеческие жесты и выполнять соответствующие команды с достаточной степенью точности.

Такие технологические гиганты как Apple, Kuka Robotics, BMW, Facebook, Netflix и другие активно развивают перспективное направление интерфейсов человеко-машинного взаимодействия, где жестовое взаимодействие является одним из самых популярных направлений. И задача точного распознавания жестовых команд является в высшей степени приоритетной.

Представленная научная работа демонстрирует систему обнаружения, отслеживания и классификации статических жестов рук в видеопотоке с использованием компьютерного зрения и методов глубокого обучения. Тема актуальна и представляет собой основу для масштабируемой системы управления на базе жестовых команд, которая может применяться в качестве интерфейса для человеко-машинного взаимодействия.

Обнаружение рук с использованием цветовой маски

Руки и тело человека имеют уникальные визуальные признаки. В задаче распознавания жестов на основе изображений, жесты состоят из фрагментов изображений рук и / или тела и использование таких фрагментов-признаков при идентификации жестов вполне резонно.

Цвет же является визуальным признаком для идентификации жестов из фоновой информации. Однако на системы распознавания жестов на основе цвета сильно влияют уровень освещения и тени [3]. Еще одна распространенная проблема обнаружения жестов по цвету кожи состоит в том, что цвета кожи у людей отличаются (рис. 1).

Рис. 1. Цветовая палитра оттенков кожи

Однако использование цветовых масок для задачи обнаружения объектов оправдано ввиду относительной простоты самого метода, особенно на стадии

прототипирования [4, 5]. Нахождение ROI (области интереса) – области кисти руки в этом случае, включает в себя следующие операции (рис. 2):

- первичное размытие изображения для удаления шумов (размытие по Гауссу);
- преобразование изображения из RGB в цветовое пространство HSV;
- определение верхней и нижней границы интенсивности пикселей HSV, которые следует рассматривать как кожу;
- обнаружение и выделение области кисти руки (ROI). Чтобы упростить решение в данном случае, ROI — это область с наибольшим количеством соседних белых пикселей;
- ROI готов к задаче классификации.

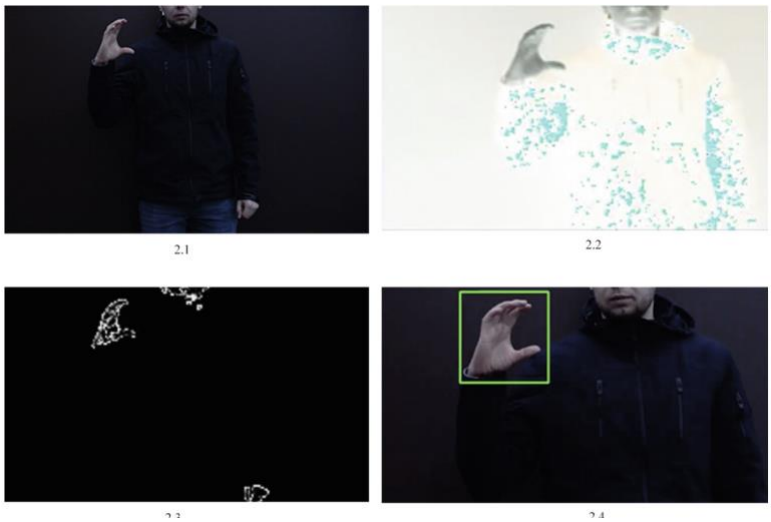

Рис. 2. Определение ROI с помощью маски цвета кожи: 2.1 – оригинал, 2.2 – фильтр HSV, 2.3 – цветовая маска, 2.4 – ROI

Набор данных и их предварительная обработка

Для обучения и тестирования классификатора был создан оригинальный набор данных, состоящий на данный момент из 2071 изображения рук в определенной жестовой конфигурации и поделенный на 10 классов. Каждый класс – это изображения жеста (киремы), обозначающего одну букву русского алфавита,

которые были выполнены при разных условиях освещения разными людьми (рис. 3).

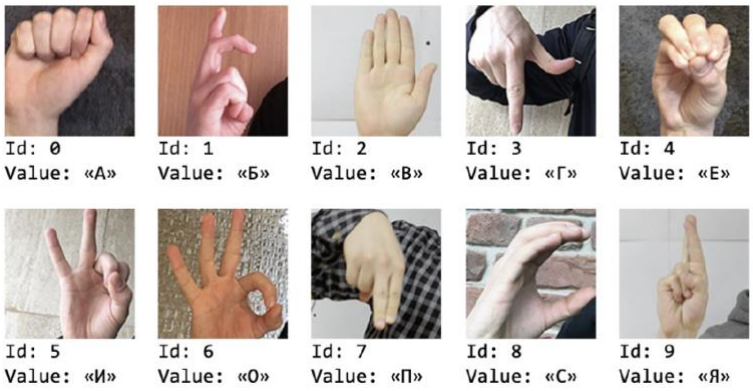

Рис. 3. Примеры из набора данных

Перед отправкой данных в классификатор необходимо выполнить предварительную обработку изображений – преобразования, которые помогут избавиться от избыточных признаков, что, в свою очередь, повысит производительность классификатора. Предварительная обработка изображений уменьшает количество избыточных деталей (например, информацию о цвете из пространства RGB) [6].

При наличии небольших наборов данных может потребоваться большой объем синтетических данных. Для этого используются методы искусственного расширения набора данных. Известно, что чем больше данных используется алгоритмом глубокого обучения, тем более эффективным он может быть. Даже когда данные низкого качества, алгоритмы могут работать лучше, если полезные признаки будут извлечены алгоритмом из исходного набора данных [7].

Следующие операции были применены, чтобы получить синтетические изображения:

– случайное вращение изображения (все направления, +/− 10 градусов);

– случайное изменение пропорций изображения.

Чтобы минимизировать переобучение классификатора, набор данных был разделен на 3 части:

- обучающая выборка содержит около 80% набора данных (1671 изображение);
- тестовая выборка содержит около 20% набора данных (400 изображений) и используется для оценки модели (эти изображения не использовались во время обучения и перекрестной проверки);
- набор для перекрестной проверки содержит около 20% обучающей выборки (300 изображений).

Архитектура нейронной сети для задачи классификации жестов и анализ ее производительности

В этом разделе описана архитектура сверточной нейронной сети, примененной для классификации жестов. В качестве эталона были взяты две модели: архитектура LeNet-5 [8] и статический классификатор жестов первого поколения [9].

Основным недостатком LeNet-5 является ее переобучение в некоторых случаях и отсутствие встроенного механизма, позволяющего минимизировать этот недостаток. LeNet-5 была улучшена за счет добавления выпадающих слоев (Dropout). Улучшенная архитектура LeNet-5 (классификатор статических жестов первого поколения) для задачи классификации жестов была представлена в статье [9]. Основным недостатком этой модели является сравнительно низкая производительность на тестовом наборе данных (91,38%).

Новая архитектура (классификатор статических жестов второго поколения) является более сложной: включает в себя дополнительные сверточные слои, а также дополнительные нейроны в полносвязных слоях (рис. 4).

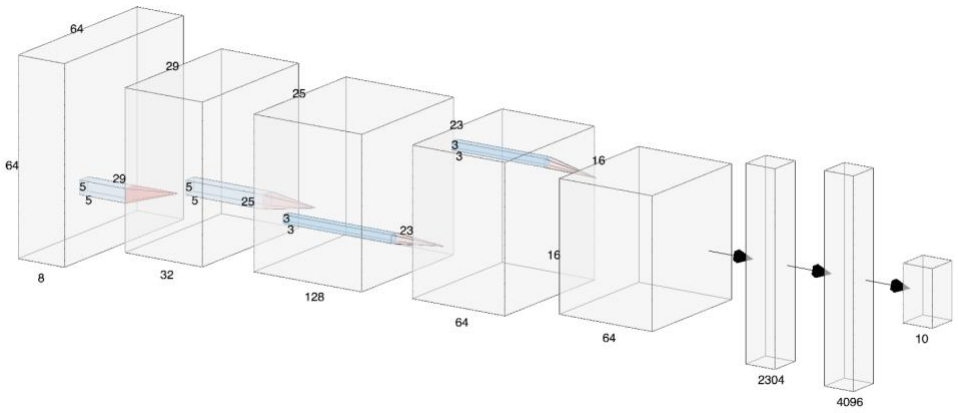

Рис. 4. Улучшенная архитектура нейронной сети для задачи классификации жестов

Процедуры обучения и тестирования были выполнены на эталонных классификаторах. Показатели производительности эталонных классификаторов и улучшенного классификатора приведены в таблице 2.

Таблица 1

Сравнение производительности эталонных классификаторов и классификатора второго поколения

Архитектура нейронной сети	**Потери (при обучении) - Loss**	**Точность (при обучении) - Accuracy**	**Потери (при тестировании) - Loss**	**Точность (при тестировании) - Accuracy**
LeNet-5	0,0681	0,9985	0,5134	0,8708
Классификатор статических жестов первого поколения [9]	0,1411	0,9369	0,2385	0,9138

Классификатор статических жестов второго поколения	0,1310	0,9433	0,2119	0,9360

Выводы

В результате исследования был разработан и опубликован уникальный набор данных из 10 классов и более 2000 уникальных изображений, а также программный комплекс для обнаружения, отслеживания и классификации статических жестов в видеопотоке с использованием компьютерного зрения и методов глубокого обучения (рис. 5).

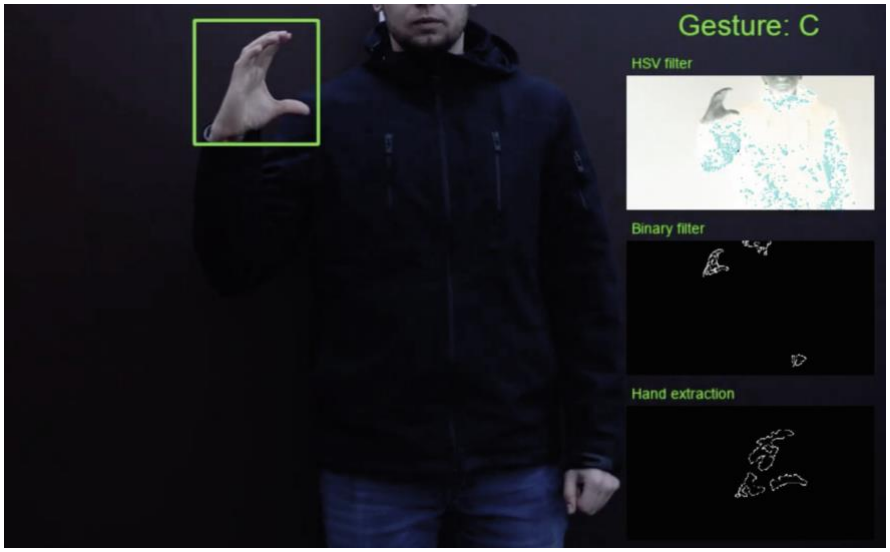

Рис. 5. Интерфейс управления жестами

Решение включает в себя модуль обнаружения кистей рук, который использует цветовую маску, модуль отслеживания жестов, модуль классификации статических жестов в обнаруженной области изображения на основе сверточной нейронной сети, а также вспомогательный модуль предварительной обработки изображений и модуль расширения набора данных. Для разработки архитектуры, обучения и тестирования нейронной сети был использован фреймворк PyTorch.

Классификатор демонстрирует точность классификации в 93,6% на тестовом наборе данных, которая выше, чем у предыдущей версии - 91,38% и у классификатора LeNet-5 - 87,08%. Представленные результаты точности являются достаточной основой для первой версии промышленного прототипа интерфейса управления жестами и дальнейших исследований в этом направлении.

Литература

[1] Kruger, J., Lien, T., Verl, A.: Cooperation of human and machines in assembly lines. CIRP Ann. Manuf. Technol – 2009, 58, 628–646.

[2] Bauer, A., Wollherr, D., Buss, M.: Human-robot collaboration: a survey. Int. J. Humanoid Rob. – 2008, 5, 47–66.

[3] Letessier, J., Berard, F.: Visual tracking of bare fingers for interactive surfaces. In: Proceedings of the 17th Annual ACM symposium on User interface software and technology - 1970, 119–122

[4] Nalepa, J., Grzejszczak, T., Kawulok, M.: Wrist Localization in Color Images for Hand Gesture Recognition. In: Gruca, D.A., Czachorski, T., Kozielski, S. (eds.) Man-Machine Interactions 3. AISC - 2014, vol. 242, 79–86.

[5] Habili, N., Lim, C., Moini, A.: Segmentation of the face and hands in sign language video sequences using color and motion cues. IEEE Trans. Circuits Syst. Video Technol. - 2004, 14(8), 1086–1097.

[6] Forstner, W.: Image preprocessing for feature extraction in digital intensity, color and range images. In: Dermanis, A., Grun, A., Sanso, F. (eds.) Geomatic Method for the Analysis of Data in the Earth Sciences. LNEARTH - 2000, vol. 95, 165–189.

[7] Wang, J., Perez, L.: The effectiveness of data augmentation in image classification using deep learning. Stanford University - 2017.

[8] LeCun, Y., Jackel, L., Bottou, L.: Learning algorithms for classification: a comparison on handwritten digit recognition. AT&T Bell Laboratories - 1995.

[9] Potkin, O., Philippovich, A.: Static gestures classification using convolutional neural networks on the example of the Russian sign language. In: Supplementary Proceedings of the Seventh International Conference on Analysis of Images, Social Networks and Texts (AIST 2018), 229–234.

HAND GESTURES DETECTION, TRACKING AND CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORK

Oleg Potkin and Andrey Philippovich
Moscow Polytechnic University
Moscow 107023, Russian Federation
olpotkin@gmail.com

Abstract. The article describes a software pipeline for detecting, tracking and classification of static hand gestures of the Russian Sign Language in a video stream using computer vision and deep learning techniques. The dataset used for this task is original, includes 10 classes and consists of more than 2000 unique images. The solution includes a hand detection module that uses a color mask, a gesture tracking module, a static gestures classification module in the detected region of the image based on convolutional neural network, as well as an auxiliary image preprocessing module and dataset augmentation module.

Keywords: Deep learning, Convolutional neural networks, Computer vision, Detection, Tracking, Classification, Hand gestures, Russian Sign Language.