Вісник Національного технічного університету "ХПІ", 2018, № 42 (1318)
ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

УДК 004.932.2 **DOI: 10.20998/2411-0558.2018.42.15**

Р. В. СІРЯК, здобув. СНУ ім. В. Даля, Сєвєродонецьк,
І. С. СКАРГА-БАНДУРОВА, д-р техн. наук, доц., зав. каф., СНУ ім. В. Даля, Сєвєродонецьк

МОДЕЛЬ ОБРОБКИ ПОТОКОВИХ ДАНИХ ДЛЯ РОЗПІЗНАВАННЯ ОКРЕМИХ ОДИНИЦЬ ЖЕСТОВОЇ МОВИ

У статті розглянута задача розпізнавання окремих жестів рук, отриманих з вебкамери. Запропоновано модель обробки потокових даних та розпізнавання жестів на відеозображеннях у вигляді 10-шарової згорткової нейронної мережі. За результатами оцінки якості моделі, отримана точність на тестовій множині склала 96%, значення функції втрати – 0.02. Результати перевірки показали, що модель є стійкою до відносно широких кутів обертання рук і є незалежною від освітлення, завдяки використанню контурів. Іл.: 3. Бібліогр.: 10 назв.

Ключові слова: модель; потокові дані; розпізнавання; жестова мова; згорткова нейронна мережа; контур.

Постановка проблеми та аналіз останніх досліджень і публікацій. Останні роки, завдяки розвитку обчислювальних технологій, з'явилися нові можливості для реалізації раніше важкоздійснюваних проектів по розпізнаванню образів. На даний момент, для розпізнавання візуальних образів найбільше використання отримали згорткові нейронні мережі (Convolutional Neuron Networks, CNN), рекурентні нейронні мережі (Recurrent Neuron Networks, RNN), мережі з довгою короткостроковою пам'яттю (Long Short-Term Memory, LSTM), їх комбінування та різні модифікації. Проте до недавнього часу для задач автоматичного розпізнавання мови жестів використовувався далеко не весь потенціал, який є сьогодні широко доступним.

Завдання розпізнаванням жестів пов'язані з великою різноманітністю проблем, що виникають кожного разу, коли необхідно розпізнати об'єкт, такий як оклюзія, зміни умов освітлення, неоднорідність та зміни фону. Оскільки, в даному випадку, задача полягає не тільки в тому, щоб знайти цільове зображення у будь який момент часу та відокремити його від фону, а також проаналізувати динамічні функції простору-часу, відстежити початок і кінець жесту в потоці наступних кадрів.

Залежно від підходів і цілей завдання розпізнавання жестів можуть вирішуватися різними методами. Так, в роботі [1] CNN використовується для вилучення ознак, а Randomized Decision Forest Classifier для сегментації зображення. В [2] використано моделі багатошарового та

багатомодального глибокого навчання. Комбінуючи дані RGBD з даними Upper-Body Skeletal Motion, CNN була успішно навчена 20 знакам італійської мови жестів. Слід, однак, зауважити, що дана техніка не призначена для використання поза приміщенням. Автори [3] використовували CNN разом з Microsoft Kinect для розпізнавання мови жестів американської англійської мови. Отримана точність алфавітних і числових знаків склала відповідно 82,5% та 97%. В роботі [4] для розпізнавання жестів була використана 3D CNN, в якій витягувалися як просторові, так і часові ознаки, фіксуючи інформацію про рух і кодуючи її в сусідніх кадрах. Автори [5] поєднали у своїй роботі 3D CNN і мультипотокову LSTM-RNN для визначення жестів та їх класифікації. В результаті такої комбінації стало легше обробляти зміни рухів. Значних успіхів здобули автори [6], використовуючи для розпізнавання італійської мови жестів CNN і Microsoft Kinect. Автори досліджували п'ять різних архітектур глибокого навчання і прийшли до висновку, що двонаправлене повторення і часова згортка можуть істотно поліпшити розпізнавання жестів. Майже всі дослідники повідомляють про отримання високих показників точності на рівні від 77,5% [7] до 97% [3, 5, 8]. На відміну від більшості реалізованих проектів, загальною метою нашого проекту є розробка методу комп'ютерного бачення, здатного розпізнавати жести української мови за допомогою смартфонів, щоб використовувати жести для взаємодії з додатком. Враховуючи вищевикладене, варто відзначити, що на даний момент, універсального підходу, який працює з високою швидкістю і точністю розпізнавання при будь-яких умовах, і може бути використовуваним в смартфонах не існує. Автори останньої публікації, присвяченої розпізнаванню жестів за допомогою смартфонів [8], наголошують на необхідності підвищення точності процесу локації. Візуальні засоби, що використовують відеокамери смартфонів дають високу точність розпізнавання, проте сильно залежать від зміни освітлення. Застосування кольорових міток допомагає усунути цю проблему, але є неприродним і незручним для повсякденного життя. Використання спеціально обладнаних рукавичок допомагає з високою точністю визначити ознаки руки, але також є незручним засобом взаємодії з комп'ютером, і, до того ж, досить дорогим. Нарешті, тривимірні сенсори, бувши абсолютно незалежними від зміни освітлення, дозволяють легко знаходити ключові ознаки і локації руки, але дають досить низькі показники розпізнавання.

Метою статті є представлення розробленої моделі для обробки потокових даних та розпізнавання жестів на відеозображеннях, як основи для створення системи розпізнаватння жестової мови за допомогою

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

смартфонів, здатної ефективно оперувати в різних середовищах відносно освітлення та кутів обертання рук.

Основна частина. Модель пропонованої згорткової нейронної мережі надано на рис. 1.

Дана мережа містить 10 шарів. До трьох згорткових шарів застосовуються відповідно 16, 32 і 64 згортальних ядра з встановленим режимом valid. Функцією активації є Rectified Linear Unit (ReLU)

$$f(x) = \max(0, x). \quad (1)$$

За кожним згортковим шаром іде шар max-pooling розміром 2 × 2 з кроком в один піксель. Після чого, дані в шарі flatten перетворюються з 2D-представлення в одновимірний вектор, і в кінці проходять через два повнозв'язних шари dense. Другий повнозв'язний шар з функцією softmax, що перетворює вектор дійсних чисел в вектор ймовірностей, є вихідним і містить softmax-класифікатор з трьома класами. Мережа навчалася з використанням категорийної функції втрат ентропії.

Для вирішення завдання розпізнавання окремих жестів створено власний набір зображень рухів рук, що був використаний для навчання та перевірки тестових наборів. Набір зображень, отримано зі звичайної веб-камери, що знімала кожен окремий жест з періодичністю 3 мс. У створюваному наборі руки мали колір шкіри європеоїда.

Світло розсіяне електричне з підстроюванням положення камери під найменший контраст між ділянками світлих і темних ділянок шкіри. Робота була виконана на двоядерному процесорі i3-7100.

Технологія розпізнавання жестів містить три основних етапи: (1) сегментація руки, (2) витяг ознак з отриманого регіону, (3) класифікація жестів.

На першому етапі отримані зображення розміром 200×200 пікселів переводилися в чорно-білий формат і піддавалися впливу фільтра Гаусова розмиття для видалення шуму [9]. Для реалізації даного фільтра використано функцію GaussianBlur бібліотеки OpenCV [10]. Після цього до них застосовувався алгоритм адаптивної порогової обробки для виділення контурів.

В даному випадку використовувалася ще одна функція бібліотеки OpenCV adaptiveThreshold.

В результаті, було отримано однотонні зображення розмірності 200 × 200 × 1, що несуть інформацію про руку, яка демонструє один з жестових знаків.

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

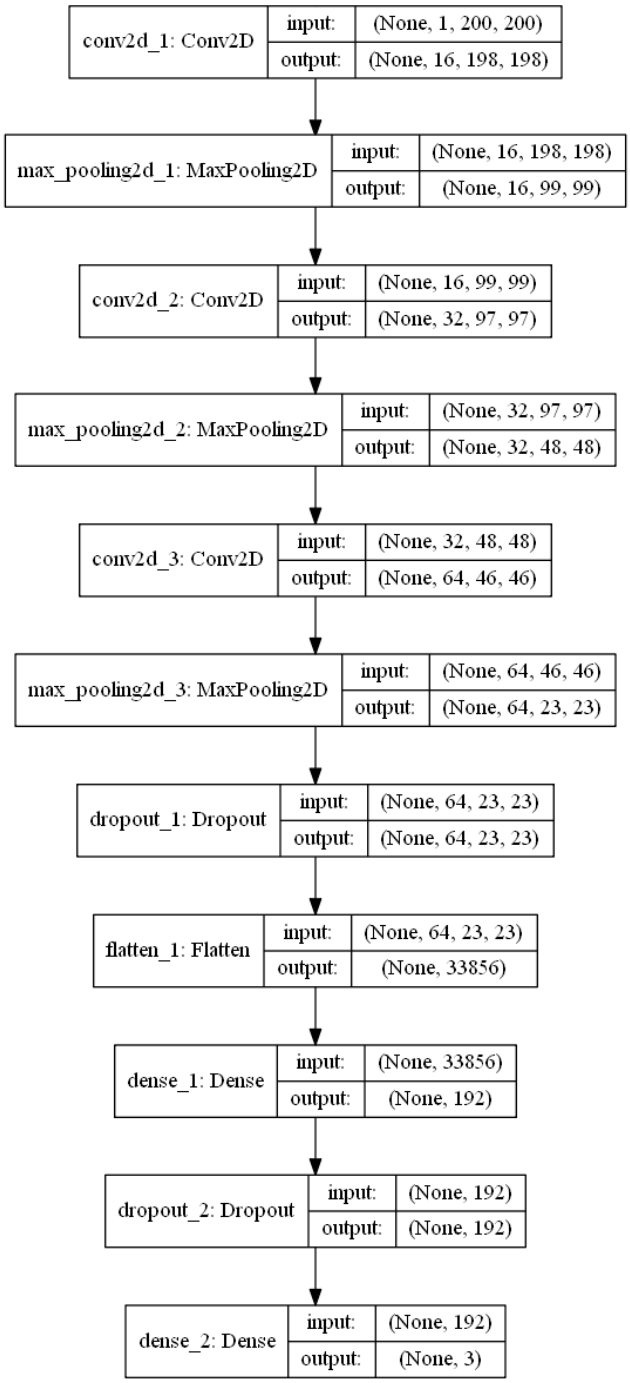

Рис. 1. Модель згорткової нейронної мережі для розпізнавання окремих одиниць жестової мови

Етап 2, витяг ознак, проводився за допомогою розробленої згорткової нейронної мережі (рис. 1). У процесі навчання нейронна мережа визначала особливості, характерні для кожного класу. Імовірність приналежності даних до класу реалізується функцією активації softmax, що перетворює вихідний сигнал останнього шару в розподіл імовірності між 0 і 1. Розмір вектора, що надходить у softmax, дорівнює кількості класів, представлених у моделі та зв'язок між функцією й розподілом імовірності рухів жестів, сформульованих як лінійна функція

$$z = W^T \cdot x + b, \tag{2}$$

де W^T позначає 2D-фільтр; $x \in R^{D\times 1}$ вхідна характеристика; $z \in R^{C\times 1}$ – змінна, що описує розподіл; C – кількість типів (класів) рухів жестів.

Значення i-го виходу в softmax визначається за формулою

$$y_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}, \tag{3}$$

де z_i – це i-й елемент z, а $y = [y_1, y_2, ..., y_c]^T$ – це вихід рівня softmax класифікатора.

На етапі класифікації, для запобігання перенавчання використовувалась функція dropout. Також, для зниження ймовірності перенавчання, штучного збільшення даних і забезпечення інваріантності класифікатора до трансформацій було застосовано Data Augmentation. Через функцію Keras fit_generator генерувалися додаткові дані з вихідного набору за допомогою афінних перетворень обертання, зрушення і зміни масштабу вихідних зображень.

В якості метода оптимізації обрано оптимізації adaptive moment estimation (adam). Adam використовує як середні значення градієнтів, так і другі імпульси градієнтів, що запобігає потраплянню в локальний мінімум. Нижче надано формулу, в якій m_t вираховує перший імпульс, а v_t – другий:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t, \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2. \end{aligned} \tag{4}$$

Результати. Для перевірки якості навчання використовувалася метрика accuracy, тобто, відношення кількості правильно передбачених значень до загальної кількості всіх відповідей. В якості функції втрати використовувалася категоріальна перехресна ентропія, тобто вираховувалася логарифмічна втрата на кілька представлених класів.

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

Якщо передбачені моделлю значення дорівнюють q, в той час як справжні значення дорівнюють p, то категоріальна перехресна ентропія буде виглядати як

$$\mathcal{L}(p, q) = -\sum_x p(x) \log (q(x)). \quad (5)$$

На рис. 2 показано зміну показника accuracy на навчальній і валідаційній вибірках протягом 50-и епох. На рис. 3 показано зміна значення функції втрат за той же період.

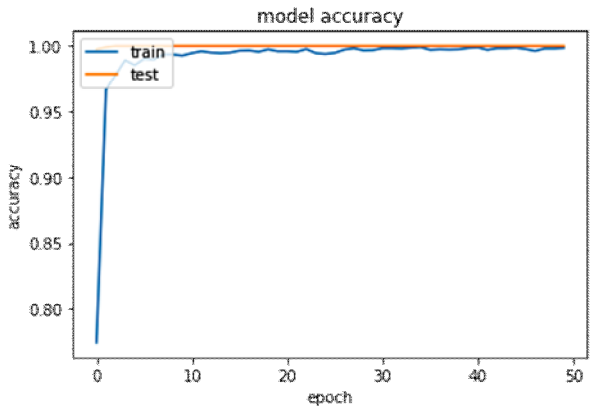

Рис. 2. Точність моделі для тренувального та тестового наборів

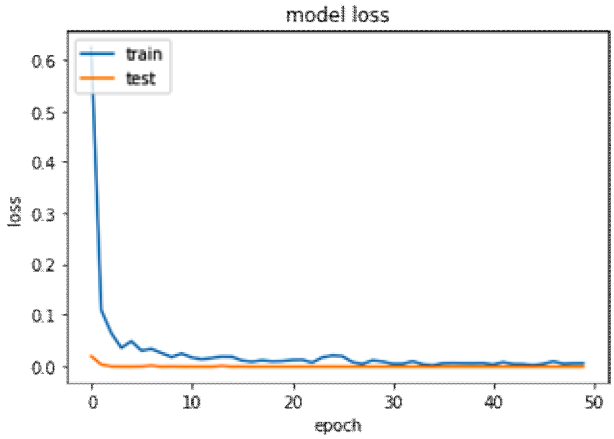

Рис. 3. Графік функції втрат для для тренувального та тестового наборів

Враховуючи той факт, що дослідження проводилися на двоядерному i3-7100 процесорі з використанням простої веб-камери отримано високий ступінь вірних прогнозів. В результаті роботи мережі досягнуто точність на тестовій множині в 96%, а значення функції втрати – 0.02.

Висновки. Розроблена модель відповідає основним принципам побудови згорткових нейронних мереж і дозволяє відстежувати і розпізнавати окремі жести у відеопотоці з високою якістю. Її точність розпізнавання з власним набором даних не гірша, ніж у відомих. Разом з тим, на відміну від існуючих, завдяки використанню контурів, модель є стійкою до відносно широких кутів обертання рук і незалежною від освітлення. При цьому, для ефективної роботи достатньо стандартної веб-камери. Серед недоліків моделі варто відзначити, що вона є не ефективною на неоднорідному зміненому фоні, і жести рук людей, які не брали участь у створенні набору даних, можуть бути гіршими. У наступному ми плануємо збільшити кількість та види впізнаваних жестів, та маємо намір поліпшити запропоновану мережу. Для розпізнавання складніших жестів до CNN буде доданий рекурентний блок. Планується розробити також засоби для поліпшення якості розпізнавання в умовах різнорідного фону.

Список літератури:

1. *Tompson J.* Real-time continuous pose recovery of human hands using convolutional networks / *J. Tompson, M. Stein, Y. LeCun, K. Perlin* // ACM Transactions on Graphics (ToG). –2014. – Vol. 33 (5). – P. 169-173.

2. *Neverova N.* Multi-scale deep learning for gesture detection and localization / *N. Neverova, C. Wolf, G.W. Taylor, F. Nebout* // Computer Vision – ECCV 2014 Workshops. ECCV. Lecture Notes in Computer Science. – 2014. – Vol. 8925. – P. 474-490.

3. *Bheda V.* Using deep convolutional networks for gesture recognition in american sign language / *V. Bheda, D. Radpour* // CoRR, abs/1710.06836. – 2017. – P. 1-5.

4. *Ji S.* 3D convolutional neural networks for human action recognition / *S. Ji, W. Xu, M. Yang, K. Yu* // IEEE transactions on pattern analysis and machine intelligence. – 2013. – Vol. 35 (1). – P. 221-231.

5. *Nishida N.* Multimodal gesture recognition using multi-stream recurrent neural network / *N. Nishida, H. Nakayama* // Image and Video Technology. PSIVT 2015. Lecture Notes in Computer Science. – 2015. – Vol. 9431. – P. 682-694.

6. *Pigou L.* Sign language recognition using convolutional neural networks / *L. Pigou, S. Dieleman, P.-J. Kindermans, B. Schrauwen* // Computer Vision – ECCV 2014 Workshops. ECCV 2014. Lecture Notes in Computer Science. – 2014. – Vol. 8925. – P. 572-578.

7. *Molchanov P.* Hand gesture recognition with 3D convolutional neural networks / *P. Molchanov, S. Gupta, K. Kim, J. Kautz* // 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). – 2015. – P. 1-7.

8. *Alashhab S.* Hand gesture detection with convolutional neural networks / *S. Alashhab, A.-J. Gallego, M.Á. Lozano* //Advances in Intelligent Systems and Computing – 2018. – P. 45-52.

9. *Сиряк Р.В.* Технологии идентификации и распознавания жестов / *Р.В. Сиряк* // Вісник Східноукраїнського національного університету ім. В. Даля. – 2017. – № 8 (238). – С. 79-85.

10. Open Source Computer Vision: Library [Електронний ресурс]. – Режим доступу: www URL: https://opencv.org/ (accesed 23.10.2018).

References:
1. Tompson, J., Stein, M., LeCun, Y. and Perlin, K. (2014), "Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks", *ACM Transactions on Graphics (ToG)*, Vol. 33 (5), pp. 169.
2. Neverova, N., Wolf, C., Taylor, G.W. and Nebout, F. (2014), "Multi-Scale Deep Learning for Gesture Detection and Localization", In: Agapito L., Bronstein M., Rother C. (eds) *Computer Vision – ECCV 2014 Workshops. ECCV. Lecture Notes in Computer Science*, Vol. 8925, pp. 474-490.
3. Bheda, V. and Radpour, D. (2017), "Using Deep Convolutional Networks for Gesture Recognition in American Sign Language", In*: CoRR, abs/1710.06836*, pp. 1-5.
4. Ji, S., Xu, W., Yang, M. and Yu, K. (2013), "3D Convolutional Neural Networks for Human Action Recognition", *IEEE transactions on pattern analysis and machine intelligence*, Vol. 35 (1), pp. 221-231.
5. Nishida, N. and Nakayama, H. (2015), "Multimodal Gesture Recognition Using Multi-Stream Recurrent Neural Network", In: Bräunl T., McCane B., Rivera M., Yu X. (eds) *Image and Video Technology. PSIVT 2015. Lecture Notes in Computer Science*, Vol 9431, pp. 682-694.
6. Pigou, L., Dieleman, S., Kindermans, P.-J., and Schrauwen, B. (2014), "Sign Language Recognition Using Convolutional Neural Networks", In: Agapito L., Bronstein M., Rother C. (eds) *Computer Vision – ECCV 2014 Workshops. ECCV 2014. Lecture Notes in Computer Science*, Vol 8925, pp. 572-578.
7. Molchanov, P., Gupta, S., Kim, K., and Kautz, J. (2015), "Hand Gesture Recognition with 3D Convolutional Neural Networks", *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1-7.
8. Alashhab, S., Gallego, A.-J., and Lozano, M.Á. (2018), "Hand Gesture Detection with Convolutional Neural Networks", *Advances in Intelligent Systems and Computing*, pp. 45-52.
9. Siryak, R.V. (2017), "Gesture Identification and Recognition Techniques", *Herald of Volodymyr Dahl East Ukrainian National University*, no. 8 (238), pp. 79-85.
10. Open Source Computer Vision Library, available at: https://opencv.org/ (accessed 23 October 2018).

Статтю представив д.т.н., проф. Національного технічного університету "Харківський політехнічний інститут" А.Е. Филатова

Надійшла (received) 16.11.2018

Siryak Rostislav, Ph.D. student
Volodymyr Dahl East Ukrainian National University
59-a Central Avenue, Severodonetsk, Luhansk region, Ukraine, 93400
tel./phone: (064) 522-89-97, e-mail: hashem.r@gmail.com
ORCID ID: 0000-0002-6775-1218

Skarga-Bandurova Inna, D.Sci.Tech., Professor
Volodymyr Dahl East Ukrainian National University
59-a Central Avenue, Severodonetsk, Luhansk region, Ukraine, 93400
tel./phone: (064) 522-89-97, e-mail: skarga_bandurova@ukr.net
ORCID ID: 0000-0003-3458-8730

УДК 004.932.2

Модель обробки потокових даних для розпізнавання окремих одиниць жестової мови / Сіряк Р.В., Скарга-Бандурова І.С. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2018. – № 42 (1318). – С. 73 – 81.

У статті розглянута задача розпізнавання жестів рук, отриманих з вебкамери. Запропоновано модель обробки потокових даних на відеозображеннях у вигляді 10-шарової згорткової нейронної мережі. За результатами оцінки якості, отримана точність на тестовій множині склала 96%, значення функції втрати 0.02. Результати перевірки показали, що модель є стійкою до відносно широких кутів обертання рук і є незалежною від освітлення, завдяки використанню контурів. Іл.: 3. Бібліогр.: 10 назв.

Ключові слова: модель, потокові дані; розпізнавання; жестова мова; згорткова нейронна мережа; контур.

УДК 004.932.2

Модель обработки потоковых данных для распознавания отдельных единиц жестового языка / Сиряк Р.В., Скарга-Бандурова И.С. // Вестник НТУ "ХПИ". Серия: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2018. – № 42 (1318). – С. 73 – 81.

В статье рассмотрена задача распознавания жестов рук, полученных с вебкамеры. Предложена модель обработки потоковых данных видеоизображения в виде 10-слойной сверточной нейронной сети. По результатам оценки качества, полученная точность на тестовом множестве составила 96%, значение функции потери 0.02. Результаты проверки показали, что модель устойчива к относительно широким углам вращения рук и малозависима от освещения. Ил.: 3. Библиогр.: 10 назв.

Ключевые слова: модель; потоковые данные; распознавание; жестовый язык; сверточная нейронная сеть; контур.

UDC 004.932.2

A model for processing stream data for the recognition of individual units of the sign language / Siryak R.V., Skarga-Bandurova I.S. // Herald of the National Technical University "KhPI". Series of "Informatics and Modeling". – Kharkov: NTU "KhPI". – 2018. – № 42 (1318). – P. 73 – 81.

The paper deals with the problem of recognizing the single hand gestures received from a webcam. The model for processing of stream data and recognition of gestures from video images in the form of a 10-layer convolutional neural network is proposed. As a result of the model quality evaluation, the accuracy obtained on the test set was 96%, the value of the loss function 0.02. The results of the test showed that the model is resistant to relatively wide angles of hand rotation and is independent of light, due to the use of contours. Figs.: 3. Refs.: 10 titles.

Keywords: model; data stream; recognition; sign language; convolutional neural network; contour.