СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВОЛОДИМИРА ДАЛЯ
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Кваліфікаційна наукова
праця на правах рукопису

СІРЯК РОСТИСЛАВ ВІКТОРОВИЧ

УДК 004.05+004.93

Д И С Е Р Т А Ц І Я

МОДЕЛІ ТА МЕТОД ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ЛЮДИНО-МАШИННОЇ ВЗАЄМОДІЇ З ВИКОРИСТАННЯМ ЖЕСТІВ

05.13.06 – інформаційні технології

12 - інформаційні технології
галузь знань

Подається на здобуття наукового ступеня кандидата технічних наук

Дисертація містить результати власних проваджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

_____ Сіряк Ростислав Вікторович
підпис

Науковий керівник Скарга-Бандурова Інна Сергіївна,
доктор технічних наук, професор

Всі примірники дисертації ідентичні за змістом.
Вчений секретар спеціалізованої
Вченої ради К 29.051.16 /Сафонова С.О./

Сєвєродонецьк – 2021

АНОТАЦІЯ

Сіряк Ростислав Вікторович. Моделі та метод інформаційної технології людино-машинної взаємодії з використанням жестів. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 “Інформаційні технології”. – Східноукраїнський національний університет імені Володимира Даля, Сєвєродонецьк, 2021.

Дисертацію присвячено вирішенню актуального науково-технічного завдання розроблення моделей та методів інформаційної технології обробки і розпізнавання статичних і динамічних жестів рук на зображеннях та відеопослідовностях у режимі реального часу. Об'єктом дослідження є системи комп'ютерного зору, що здійснюють ідентифікацію та класифікацію жестів рук або інших об'єктів на зображенні і у відеопослідовності, а предметом дослідження – математичні моделі, обробка даних і алгоритми розпізнавання статичних і динамічних жестів руки для вирішення задач людино-машинної взаємодії.

Виконано аналіз літературних джерел за темою дисертації, проаналізовано особливості існуючих підходів до вирішення завдань розпізнавання, захвату і відстеження жестів руки. Встановлено, що використання методів глибокого навчання і нейромережевого програмування дає найкращі результати в розпізнаванні візуальних образів.

За результатами проведеного аналізу сформульоване основне науково-прикладне завдання роботи, яке полягає в розробці моделей і методів інформаційної технології для системи розпізнавання статичних та динамічних жестів для покращення людино-машинної взаємодії з використанням жестів. Визначені задачі досліджень, обрано стратегію та обґрунтовано методику й математичний апарат досліджень.

Математичний апарат дослідження базується на методах загальної теорії систем, системного аналізу, методах структурного синтезу, аналізу та

моделювання процесів, технології збільшення даних (data augmentation), технології попередньої обробки даних (очищення, фільтрація розмиття за Гауссом, перетворення кольорів, адаптивна порогова обробка, алгоритм Otsu, детектор контурів Canny, модель дифузного освітлення, моделі деформації зображень, порівняння зображень із спотвореннями), методи зворотного поширення похибки та спряжених градієнтів, методи статичної та динамічної кластеризації, методи математичної статистики, функції просторового домену, теорія автоматів, теорія подібності.

Вперше розроблено методологію розробки і спільного використання візуальних методів розпізнавання жестів рук, яка дозволяє створювати і досліджувати моделі глибокого навчання для розпізнавання статичних та динамічних жестів, здатних працювати в режимі реального часу і забезпечує розуміння способів їх налаштування для різних інтерфейсів управління жестами та потенційних застосувань в системах ЛМВ;

Удосконалено модель статичного розпізнавання жестів, побудовану згідно пропонованої методології на основі згорткової нейронної мережі, шляхом штучного збільшення даних і використання контурів. Завдяки використанню контурів, модель є стійкою до відносно широких кутів обертання рук і незалежною від освітлення. Модель із доповненими даними досягла точності 97,12%, що майже на 4% перевищує модель без доповнень (92,87%), при цьому, для ефективної роботи достатньо стандартної веб-камери.

Дістала подальшого розвитку технологія розпізнавання та прогнозування жестів на основі моделі генерації послідовностей з використанням ConvLSTM2D і Conv3D. Для подальшого розширення можливостей ConvLSTM для вирішення проблеми прогнозування жестів було виконано наступні модифікації: (1) на етапі попередньої обробки введено виявлення контура, реалізованного у вигляді фільтру. Фільтрація на вхідному рівні дозволила фіксувати часові залежності та зменшити шум вхідного сигналу; (2) проведена заміна 3 шарів batch normalization на 3 шари dropout regularization що дозволило знизити перенавчання та мінімізувати помилки передбачення при прогнозуванні

жестів. Отримана точність склала 90%, що на 30% краще ніж у моделі ConvLSTM з batch normalization (60%).

Удосконалено модель скінченного автомату для безконтактного управління переглядом медичних зображень за допомогою жестів яка, на відміну від існуючих, використовує дані прогнозованих кадрів відеопослідовностей, що дозволяє зменшити час відгуку системи.

Створено *новий датасет* для тестування пропонованих моделей і методу інформаційної технології розв'язування задач розпізнавання та прогнозування жестів в операційній залі і виконано концептуальну розробку інтуїтивного словникового запасу динамічних жестів, що дозволяє реалізувати ефективну безконтактну інтерактивну систему, адаптовану до особливостей хірургічного контексту.

Удосконалено структурну модель інформаційної технології ЛМВ з використанням жестів, за рахунок визначення основних етапів та інформаційних потоків створення та інтеграції моделей глибокого навчання, яка забезпечує прийняття рішень щодо застосування розроблених методів, засобів і технологій.

Усі теоретичні розробки дисертації доведено до конкретних інженерних методик та алгоритмів з застосуванням запропонованої інформаційної технології ідентифікації та класифікації статичних і динамічних жестів рук на основі глибоких нейронних мереж. В результаті виконаного дисертаційного дослідження розроблено програмні реалізації моделей системи розпізнавання, яка демонструє високу точність і швидкодію з низкою чутливістю до умов освітлення. Результати дисертаційної роботи використовуються в якості базових елементів системи ЛМВ для навігації та перегляду медичних зображень в операційній залі. Розроблені моделі, методи, інформаційна технологія використовувались при розробці проекту клінічної системи безконтактного перегляду медичних зображень медичного центру "Твоя лікарня", м. Сєвєродонецьк.

Ключові слова: людино-машинна взаємодія, жести рук, глибоке навчання

Список публікацій здобувача за темою дисертації

Праці, які відображають основні наукові результати дисертації

1. Skarga-Bandurova I. Siriak R., Bilodorodova T., Cuzzolin F., Bawa V.S., Mohamed M.I., Charles R.D. “Surgical Hand Gesture Prediction for the Operating Room”, *Studies in Health Technology and Informatics* (Eds. Bernd Blobel, Lenka Lhotska, Peter Pharow, Filipe Sousa). 2020, Vol. 273, pp. 97-103. (Індексується в SCOPUS).

2. Сіряк Р.В., Скарга-Бандурова І.С., Шумова Л.О. “Особливості технології обробки даних для розпізнавання жестів”, *Вісник Національного технічного університету "ХПІ". Збірник наукових праць. Серія: Інформатика та моделювання*. Харків: НТУ "ХПІ". 2019, №.13(1338). С. 112-122.

3. Білобородова Т.О., Сіряк Р.В., Скарга-Бандурова І.С., Давіденко М.О., Сіроштан І.В., Приймак С.О. “Розпізнавання взіємодії інструментів з тканиною на медичних відеозображеннях”, *Наукові вісті Далівського університету : електронне наукове фахове видання.* 2019, № 17. Режим доступу: http://nvdu.snu.edu.ua/wp-content/uploads/2020/03/2019_17_5.pdf

4. Сіряк Р.В., Скарга-Бандурова І.С., Білобородова Т.О. Towards an empirical hyper parameters optimization in CNN”, *Вісник Східноукраїнського національного університету ім. В. Даля*. Сєвєродонецьк: СНУ ім. В. Даля. 2019, № 5(253). С. 87-91.

5. Сіряк Р.В., Скарга-Бандурова І.С. “Модель обробки потокових даних для розпізнавання окремих одиниць жестової мови”, *Вісник Національного технічного університету "ХПІ". Збірник наукових праць. Серія: Інформатика та моделювання*. Харків: НТУ "ХПІ". 2018, № 42(1318). С. 73-81.

6. Болтов Є.В., Сіряк Р.В., Щербаков Є.В., Лифар О.К. “Тестування продуктивності операційних систем реального часу для комп’ютерного зору і обробки зображень на одноплатних комп’ютерах”, *Вісник Східноукраїнського національного університету ім. В. Даля*, Сєвєродонецьк: СНУ ім. В. Даля. 2018, № 6(247). С. 23-26.

7. Сіряк Р.В. “Технологии идентификации и распознавания жестов”, *Вісник Східноукраїнського національного університету ім. В. Даля*, Сєвєродонецьк: СНУ ім. В. Даля. 2017, № 8(238). С.79-85.

Наукові праці, які засвідчують апробацію матеріалів дисертації

8. Siriak R., Skarga- Bandurova I., Boltov Y. “Deep Convolutional Network with Long Short-Term Memory Layers for Dynamic Gesture Recognition”, *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2019 10th IEEE International Conference*, 2019. Vol. 1, pp. 158-162. (Індексується в SCOPUS).

9. Boltov Y., Hrushka M., Kotsiuba I., Skarga-Bandurova I., Krivoulya G., Siriak R. “Performance Evaluation of Real-Time System for Vision-Based Navigation of Small Autonomous Mobile Robots”, *2019 IEEE 10th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, UK, Leeds, June 5-7, 2019. pp. 218-222. (Індексується в SCOPUS).

10. Сіряк Р.В., Скарга-Бандурова І.С. Система розпізнавання динамічних жестів за допомогою мережі CNN-LSTM. *Комп’ютерні інтелектуальні системи та мережі. Матеріали XI Всеукраїнської науково практичної WEB конференції аспірантів, студентів та молодих вчених* (21-23 березня 2019 р.). – Кривий Ріг: ДВНЗ «Криворізький національний університет», 2019. – С. 94-95.

11. Siriak R., Skarga-Bandurova I., Biloborodova T. Hyperparameters Optimization in CNN for Image Recognition. *Theoretical and Applied Computer Science and Information Technology: Proceedings of the III International Conference TACSIT-2019,* May 8-9 2019. Severodonetsk: Volodymyr Dahl East Ukrainian National University, pp. 21-22.

12. Сіряк Р.В., Скарга-Бандурова І.С. Аналіз ефективності адаптивних методів оптимізації гіперпараметрів згорткових нейронних мереж. *Проблеми інформатики і моделювання. Тезиси дев’ятнадцятої міжнар. наук.-техн. конф*. – Харків: НТУ "ХПІ", 2019. С. 76.

13. Сіряк Р.В., Скарга-Бандурова І.С. Використання конволюційної нейронної мережі для розпізнавання динамічних образів. *Сучасні технології в освіті та науці: Матеріали міжнар. наук.-практ. конф.*; 19 – 22 лютого 2018 р., м. Сєвєродонецьк / гол. ред. О.І. Рязанцев – Сєвєродонецьк: вид-во СНУ ім. В. Даля, 2018. С. 45-47.

14. Сіряк Р.В., Скарга-Бандурова І.С. Вибір архітектури глибинного навчання для системи розпізнавання жестів рук. *Інформаційні технології: наука, техніка, технологія, освіта, здоров'я: тези доповідей XXVI міжнародної науково-практичної конференції MicroCAD-2018*, 16-18 травня 2018р.: у 4 ч. Ч. IV. / за ред. проф. Сокола Є.І. – Харків: НТУ «ХПІ». С. 203.

15. Сіряк Р.В. Средства создания отчетной медицинской документации в медицинской информационной системе. *Збірник тез доповідей міжнародної конференції студентів, аспірантів та молодих вчених "Технологія-2012"*, Сєвєродонецьк: Технол. ін-т Східноукр. Нац. ун-ту ім. В. Даля (м. Сєвєродонецьк), 2012. С. 94-95.

16. Сіряк Р.В., Висоцький Д.В. Система розпізнавання жестів рук через реалізацію моделі нейронної мережі. *IT-Ідея – 2018: збірник науково-практичних праць*. Сєвєродонецьк : Вид-во Східноукр. ун-ту ім. В. Даля, 2018. С. 80-81.

ABSTRACT

Rostyslav Siriak, Models and Information Technology for Human-Machine Interaction Using Hand Gestures – Manuscript copyright.

The dissertation on competition of a scientific degree of the candidate of technical sciences on a specialty 05.13.06 "Information technology". - Volodymyr Dahl East Ukrainian National University, Severodonetsk, 2021.

The thesis is devoted to solving the actual scientific and applied problem of developing models and the method of information technology of processing and recognition of static and dynamic hand gestures on images and video sequences in real time. The object of the research is the computer vision systems that identify and classify

hand gestures or other objects in an image and video sequence. The subject of the research is the mathematical models, data processing and algorithms for recognition of static and dynamic hand gestures to solve problems of human-machine interaction.

The paper analyzes the literature on the topic of the thesis, analyzes the features of the existing approaches to the decision of problems of recognition, capture and tracking of hand gestures. It is established that the use of deep learning methods and neural network programming gives the best results in the recognition of visual images.

According to the results of the analysis, the main applied research task was formulated, which is to develop models and methods of information technology to develop models and methods of information technology for the system of recognition of static and dynamic gestures to improve human-machine interaction using gestures. The tasks of research are determined, the strategy is determined, and the methodology and mathematical apparatus of research are grounded.

Selected research methods are based on on methods of general systems theory, systems analysis, methods of structural synthesis, analysis and modeling of processes, data augmentation technologies, data pre-processing technologies (cleaning, Gaussian blur filter, color transformation, adaptive threshold processing, Otsu algorithm, Canny edge detector, diffuse illumination models, image deformation models, comparison of images with distortions), methods of backpropagation and conjugate gradients, methods of static and dynamic clustering, methods of mathematical statistics, spatial domain functions, automata theory, similarity theory.

For the first time, a methodology for developing and sharing visual hand gesture recognition techniques was developed to create and explore deep learning models for recognizing of static and dynamic gestures that are capable to work in real time and provides an understanding of how to configure them for different gesture control interfaces and potential applications in HMI systems.

The model of static gesture recognition, built according to the proposed methodology and based on a convolutional neural network, with image augmentation and gesture image filtering, was improved. Thanks to using of contours, the model is resistant to relatively wide angles of rotation of the hands and independent of lighting.

The model with augmented data reached an accuracy of 97.12%, which is almost 4% higher than the model without augmentation (92.87%), while a standard webcam is enough for effective performance.

Gesture recognition and prediction technology based on the sequence generation model using ConvLSTM2D and Conv3D was further developed. To further expanding the capabilities of ConvLSTM to address the issue of gesture prediction, the following modifications were made: (1) at the stage of pre-processing was introduced the detection of the contour, implemented in the form of a filter. Filtering at the input level allowed to record time dependences and reduce the noise of the input signal; (2) 3 layers of batch normalization were replaced by 3 layers of dropout regularization, which allowed to reduce overfitting and minimize prediction errors. The resulting accuracy was 90%, which is 30% better than the model of ConvLSTM with batch normalization (60%).

The model of the finite-state machine for contactless control of viewing of medical images with gestures is improved. Unlike existing models, it uses data from predicted frames of video sequences, which reduces system response time.

A new dataset was created to test the proposed models and methods of information technology for solving problems of gesture recognition and prediction in the operating room and conceptual development of an intuitive vocabulary of dynamic gestures, which allows to implement an effective contactless interactive system adapted to the surgical context.

The structural model of HMI information technology with the use of gestures was improved by identifying the main stages and information flows of creating and integrating deep learning models, which provides decision-making on the application of developed methods, tools and technologies.

All the theoretical development of the thesis brought to the specific engineering techniques and algorithms using the proposed models, methods, information technology of identification and classification of static and dynamic hand gestures based on deep neural networks. In light of this research, software implementations of models for static and dynamic gesture recognition system were developed, which

demonstrates high accuracy and speed with low sensitivity to lighting conditions. The results of the thesis are used as the core elements of the HMI system for navigation and viewing of medical images in the operating room.

The developed models, methods, information technology were used in the development of the project in the clinical system of contactless control of medical images of the medical center "Tvoya Likarnya", Severodonetsk, Ukraine.

Keywords: human-machine interaction, hand gesture recognition, deep learning

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

ЕПР	експерти, що приймають рішення	
ІКТ	інформаційно-комунікаційні технології	
ЛМВ	людино-машинна взаємодія	
МН	машинне навчання	
НМ	нейрона мережа	
ОЗ	операційна зала	
CMOS	Complementary metal oxide semiconductor sensors	датчики зображення типу “метал--оксид-напівпровідник”
CNN	Convolutional Neural Network	згорткова нейронна мережа
CTC	Connectionless Temporal Classification	тимчасова класифікація без зв’язку
IoU	Intersection over Union	перетин над з'єднанням
FC	Fully Connected	повноз'єднаний (шар)
FSM	Finite-state machine	скінченний автомат
HCI	Human-Computer Interaction	взаємодія людина-комп’ютер
HMI	Human-Machine Interface	людино-машинний інтерфейс
HMM	Hidden Markov Model	прихована модель Маркова
HOG	Histogram of Oriented Gradient	гісторгама орієнтованого градієнта
LSTM	Long Short-Term Memory	довга короткочасна пам’ять
LRCN	Long-Term Recurrent Convolutional Network	довгострокова періодична згорткова мережа
MLP	Multi-Layer Perceptron	багатошаровий перцептрон
ReLU	Rectified Linear Unit	блок лінійної ректифікації
RF	Random Forest	випадковий ліс
RNN	Recurrent Neural Network	рекурентна нейронна мережа
SVM	Support-Vector Machines	метод опорних векторів

ВСТУП

Обґрунтування вибору теми дослідження. Розпізнавання жестів рук на основі комп’ютерного зору є активною сферою досліджень людино-машинної взаємодії (ЛМВ) оскільки жести є природним засобом спілкування людей між собою, а нещодавно і з пристроями в інтелектуальному середовищі. Інтелектуальні системи розпізнавання жестів відкривають нову еру взаємодії людина-комп’ютер, тенденція ЛМВ рухається до відстеження та розпізнавання жестів рук у режимі реального часу для використання у різноманітних сферах застосування. Усвідомлюючи можливості безконтактної взаємодії, ми знаходимось у дуже захоплюючий час розвитку цих технологій. Численні ініціативи у всьому світі свідчать про надзвичайний прогрес, досягнутий за останні роки. Разом з тим, останні дослідження також висвітлюють різноманітні способи, за допомогою яких можливо підійти до цього проблемного простору. Інтерфейси, засновані на сприйнятті жестів та комп'ютерному зорі, повинні відповідати вимогам взаємодії на основі щоденної діяльності й підтримувати неформальну та неструктуровану інформацію. Останнє означає відсутність потреби мати при собі спеціальне обладнання для реалізації ЛМВ. На даний момент мета полягає не лише в демонстрації можливостей безконтактного контролю за допомогою систем комп’ютерного зору, але й у важливих технологічних та дизайнерських викликах щодо конкретних способів досягнення ЛМВ, починаючи від словникового запасу використовуваних жестів, умов та способів введення жестових команд та закінчуючи конкретними технічними проблемами вибору, проектування та використання механізмів їх розпізнавання. Насьогодні, методи машинного навчання, зокрема глибоке навчання, стали золотим стандартом у сфері комп'ютерного зору.

Питаннями розробки нових моделей і методів машинного навчання для розпізнавання жестів рук займалися S. Alashhab, V. Bheda, C. Cadoz, G. Devineau, J. Donahue, S. Hussain, P. Molchanov, N. Neverova, N. Nishida, L. Pigou, K. Simonyan, N. Tavari, та ін. В напрямах розпізнавання жестів рук, зокрема напрямі

розвитку інформаційних технологій моделювання, перекладу та навчання української жестової мови активно працюють провідні вітчизняні вчені О.В. Бармак, М.В. Давидов, Ю.В. Крак, Ю.П. Кондратенко, В.В. Пасічник, І.В. Сергієнко та ін.

За останні кілька десятиліть розроблені різні інструменти і алгоритми для реалізації людино-машинної взаємодії, для керування роботами, відеоспостереження, комунікації, іграх та інших областях, де рухи, що здійснюються людиною можуть бути інтерпретовані як певні команди або елементи комунікаційної взаємодії.

Незважаючи на широкі можливості застосування технологій, в галузі розпізнавання жестів рук існує цілий ряд невирішених завдань. Так, існуючі техніки показують досить високі показники, але переважно в спеціально створених для цього умовах. У реальному житті їх продуктивність значно знижується внаслідок різних умов освітлення, шумів зображення, неоднорідності фону, оклюзій і руху камери, тощо. Крім цього, в більшості систем невирішеною залишається проблема визначення положення руки по відношенню до камери – рука повинна знаходиться в обумовленому місці і не ухилятися далі деякого кута. Проблемою також залишається визначення початку і кінця жесту в послідовності кадрів. Все це обумовлює необхідність проведення подальших досліджень, спрямованих на розробку комплексних систем, що використовують сильні сторони різних підходів, взаємно перекривають недоліки. Існує широкий вибір алгоритмів оптимізації та обчислювання втрат, однак знаходження оптимального шляху до цільової точки потребує подальшого покращення, що в кінцевому підсумку підвищить ефективність обробки даних та результативність розпізнавання образів у цілому.

Таким чином, *актуальним науковим завданням* є розроблення моделей та методу інформаційної технології обробки і розпізнавання статичних і динамічних жестів руки на зображеннях та відеопослідовностях у режимі реального часу для покращення людино-машинної взаємодії з використанням жестів.

Зв'язок роботи з науковими програмами, планами, темами. Дослідження, результати яких викладені в дисертації, проводилися здобувачем на кафедрі комп'ютерних наук та інженерії Східноукраїнського національного університету імені Володимира Даля відповідно до державних програм і планів НДР, а також міжнародних проектів і програм: міжнародного проекту Європейського Союзу ERASMUS+ "Internet of Things: Emerging Curriculum for Industry and Human Applications", ALIOT, реєстр. номер 573818-EPP-1-2016-1-UK-EPPKA2-CBHE-JP, 10.2016-10.2020 рр; НДР "Система управління медичною інформацією" (Східноукраїнський національний університет імені Володимира Даля, номер ДР 0111U001748, 2011-2015 рр.); НДР "Дослідження методів аналізу даних в медицині" (Східноукраїнський національний університет імені Володимира Даля, номер ДР 0116U008700, 10.2016-08.2020 рр.), "Інтегрована система віддаленого моніторингу стану здоров'я" (Східноукраїнський національний університет імені Володимира Даля, номер ДР 0120U102758, 06.2020-12.2022 рр.), де здобувач брав участь як виконавець окремих розділів.

Роль автора у зазначених науково-дослідних роботах і проектах, у яких дисертант був безпосереднім виконавцем, полягає в розробленні і реалізації моделей, методів та алгоритмів розпізнавання візуальних образів з використанням нейромережевих архітектур.

Метою дисертації є покращення характеристик автоматичного розпізнавання статичних та динамічних жестів рук за рахунок розробки та практичного використання моделей і методу інформаційної технології людино-машинної взаємодії з використанням жестів.

Для досягнення поставленої мети розв'язуються такі **задачі:**

1) провести аналіз предметної області, вимог і передумов до розробки і використання систем ЛМВ за допомогою жестів;

2) розробити методологію проектування і спільного використання візуальних методів розпізнавання жестів рук;

3) розробити моделі комп'ютерного зору на основі глибоких нейронних мереж, що забезпечують відстеження та розпізнавання заданих жестів рук,

здатних функціонувати в режимі реального часу, інваріантних до афінних перетворень і зміни освітлення;

4) удосконалити модель статичного розпізнавання жестів;

5) удосконалити технологію розпізнавання та прогнозування динамічних жестів на основі моделі генерації послідовностей з використанням ConvLSTM;

6) створити новий датасет і виконати концептуальну розробку інтуїтивного словникового запасу динамічних жестів, адаптованих до особливостей хірургічного контексту;

7) удосконалити технологію взаємодії між хірургом та системою перегляду зображень в операційній залі яка використовує функції просторового домену для відстеження потенційних напрямків руху рук за допомогою скінченного автомату;

8) розробити елементи інформаційної технології для проектування системи ЛМВ з використанням жестів;

9) виконати тестування та практичне впровадження розроблених моделей, засобів і технологій.

Об'єкт дослідження – процеси людино-машинної взаємодії, що здійснюють ідентифікацію, відстеження та класифікацію жестів рук або інших об'єктів на зображенні та у відеопослідовності.

Предмет дослідження – моделі та метод інформаційної технології ідентифікації та класифікації статичних і динамічних жестів рук для вирішення задач людино-машинної взаємодії у реальному часу.

Методи дослідження. Проведені в роботі дослідження ґрунтуються на наступних методах і технологічних підходах:

- для аналізу предметної області та вимог до розробки і використання систем ЛМВ за допомогою жестів використано пошук в академічних БД, патентний пошук, технології окреслення, коментування, конспектування, схематизація, анатомування, конденсування, аналіз, синтез, агрегування;

- при розробленні методології проектування і спільного використання візуальних методів розпізнавання жестів рук використано методи загальної

теорії систем, системного аналізу, методи структурного синтезу, аналізу та моделювання процесів;

- при розробленні моделей комп'ютерного зору на основі глибоких нейронних мереж застосовано методи машинного навчання, технології збільшення даних (data augmentation), технології попередньої обробки даних (очищення, фільтрація розмиття за Гаусом, перетворення кольорів, адаптивна порогова обробка, алгоритм Otsu, детектор контурів Canny, для відділення зон зображення, які містять обличчя та долоні, використано модель дифузного освітлення, для визначення форми долоні використані моделі деформації зображень, порівняння зображень із спотвореннями);

- при розробленні технології розпізнавання та прогнозування динамічних жестів використано моделі генерації послідовностей з використанням ConvLSTM, для навчання нейромережевого класифікатора використано методи зворотного поширення похибки та спряжених градієнтів, для стеження за долонями використані методи статичної та динамічної кластеризації;

- при розробленні технології взаємодії між хірургом та системою перегляду зображень в операційній залі використано методи математичної статистики, функції просторового домену, теорія автоматів, теорія подібності.

Наукова новизна результатів досліджень полягає у наступному.

1) *вперше розроблено* методологію розробки і спільного використання візуальних методів розпізнавання жестів рук, яка дозволяє створювати і досліджувати моделі глибокого навчання для розпізнавання статичних та динамічних жестів, здатних працювати в режимі реального часу і забезпечує розуміння способів їх налаштування для різних інтерфейсів управління жестами та потенційних застосувань в системах ЛМВ;

2) *удосконалено* модель статичного розпізнавання жестів, побудовану згідно пропонованої методології на основі згорткової нейронної мережі, шляхом штучного збільшення даних і використання контурів. Завдяки використанню контурів, модель є стійкою до відносно широких кутів обертання рук і незалежною від освітлення. Модель із доповненими даними досягла точності

97,12%, що майже на 4% перевищує модель без доповнень (92,87%), при цьому, для ефективної роботи достатньо стандартної веб-камери;

3) *дістала подальшого розвитку* технологія розпізнавання та прогнозування жестів на основі моделі генерації послідовностей з використанням ConvLSTM2D і Conv3D. Для подальшого розширення можливостей ConvLSTM для вирішення проблеми прогнозування жестів було виконано наступні модифікації: (1) на етапі попередньої обробки введено виявлення контура, реалізованного у вигляді фільтру. Фільтрація на вхідному рівні дозволила фіксувати часові залежності та зменшити шум вхідного сигналу; (2) проведена заміна 3 шарів batch normalization на 3 шари dropout regularization що дозволило знизити перенавчання та мінімізувати помилки передбачення при прогнозуванні жестів. Отримана точність склала 90%, що на 30% краще ніж у моделі ConvLSTM з batch normalization (60%);

4) *удосконалено* модель скінченного автомату для безконтактного управління переглядом медичних зображень за допомогою жестів яка, на відміну від існуючих, використовує дані прогнозованих кадрів відеопослідовностей, що дозволяє зменшити час відгуку системи;

5) для тестування пропонованих моделей і методу інформаційної технології розв'язування задач розпізнавання та прогнозування жестів в операційній залі створено *новий датасет* і виконано концептуальну розробку інтуїтивного словникового запасу динамічних жестів, що дозволяє реалізувати ефективну безконтактну інтерактивну систему, адаптовану до особливостей хірургічного контексту;

6) *удосконалено* структурну модель інформаційної технології ЛМВ з використанням жестів, за рахунок визначення основних етапів та інформаційних потоків створення та інтеграції моделей глибокого навчання, яка забезпечує прийняття рішень щодо застосування розроблених методів, засобів і технологій.

Практичне значення одержаних результатів полягає в тому, що розроблені методологія, моделі та метод створюють концептуальну основу для розробки інформаційної технології людино-машинної взаємодії з використанням

жестів в режимі реального часу. Основні наукові положення дисертації реалізовано в вигляді розрахункових моделей, інженерних алгоритмів, а також відповідних програмних засобів, які утворюють прикладну інформаційну технологію ідентифікації та класифікації статичних і динамічних жестів рук на основі глибоких нейронних мереж.

В результаті виконаного дисертаційного дослідження розроблено програмні реалізації моделей системи розпізнавання, яка демонструє високу точність і швидкодію з низкою чутливістю до умов освітлення. Результати дисертаційної роботи використовуються в якості базових елементів системи ЛМВ для навігації та перегляду медичних зображень в операційній залі.

Реалізація результатів та впровадження. Результати дисертаційної роботи впроваджено

– у Східноукраїнському національному університеті імені Володимира Даля при розробці навчальних курсів “Вступ до інтерактивного проектування”, “Машинне навчання” на кафедрі комп’ютерних наук та інженерії, виконанні міжнародного проекту ERASMUS+ ALIOT 573818-EPP-1-2016-1-UK-EPPKA2-CBHE-JP “Internet of Things: Emerging Curriculum for Industry and Human Applications” (акт впровадження від 04.01.2021 р.);

– у медичному центрі “Твоя лікарня” при розробці проекту клінічної системи безконтактного перегляду медичних зображень, м. Сєвєродонецьк (акт впровадження від 21.11.2019 р.);

Особистий внесок здобувача полягає в розробці нових і удосконаленні існуючих моделей та інструментальних засобів, які забезпечують вирішення поставлених у дисертації завдань. Всі положення, які виносяться на захист, належать автору особисто. Роботи [7, 15] опубліковані без співавторів. У роботах, опублікованих у співавторстві, здобувачеві належать такі результати: результати аналізу вимог та моделювання систем нейромережевої архітектури [2, 6, 14], розроблені інструментальні засоби для автоматизації процесів автоматичного розпізнавання статичних та динамічних жестів рук [3, 9], аналіз критеріїв оптимальності для якісної оцінки вихідних даних моделі [4, 11, 12],

модель статичного розпізнавання жестів [7, 16], технологія розпізнавання та прогнозування жестів на основі моделі генерації послідовностей з використанням ConvLSTM2D і Conv3D [1], моделі розпізнавання динамічних жестів за допомогою мережі CNN-LSTM [8, 10, 13].

Апробація результатів наукових досліджень проводилася на 10 міжнародних і національних науково-практичних конференціях, а саме: 10th IEEE International Conference on Dependable Systems, Services and Technologies (DESSERT), UK, Leeds, 2019; 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Metz, France, 2019; II International Conference on Theoretical and Applied Computer Science and Information Technology (TACSIT), Сєвєродонецьк, 2019; XI Всеукраїнській науково практичної WEB конференції аспірантів, студентів та молодих вчених "Комп'ютерні інтелектуальні системи та мережі", Кривий Ріг, 2019; XIX міжнародній конференції "Проблеми інформатики і моделювання", Харків, 2019; "Сучасні технології в освіті та науці", 2018; "IT-Ідея", Сєвєродонецьк, 2018; XXVI міжнародній науково-практичній конференції "Інформаційні технології: наука, техніка, технологія, освіта, здоров'я" MicroCAD-2018, Харків, 2018; міжнародній конференції студентів, аспірантів та молодих вчених "Технологія-2012", Сєвєродонецьк, 2012.

Публікації. За темою дисертації з викладенням її основних результатів опубліковано 16 наукових праць, серед яких 6 статей у наукових фахових виданнях України [2-7], 1 стаття у фаховому виданні, індексованому у міжнародній наукометричній базі Scopus [1]; 2 публікації в працях міжнародних конференцій, які включено до бази даних Scopus [8,9], 7 публікацій за матеріалами наукових конференцій [10-16].

Структура та обсяг дисертації. Дисертація складається з анотацій, вступу, 5 розділів, висновків, додатків та списку використаних джерел зі 155 найменувань на 18 сторінках. Загальний обсяг дисертації становить 177 сторінок, з них 153 сторінки основного тексту, 40 рисунків, 5 таблиць.

РОЗДІЛ 1
АНАЛІЗ СТАНУ ПИТАННЯ РОЗВИТКУ ТА ВИКОРИСТАННЯ ТЕХНОЛОГІЙ ЛЮДИНО-МАШИННОЇ ВЗАЄМОДІЇ З ВИКОРИСТАННЯМ ЖЕСТІВ РУК

Розділ присвячений аналізу предметної області, зокрема вимог до систем людино-машинної взаємодії (ЛМВ). Представлено аналіз технологій у сфері розпізнавання та прогнозування жестів рук, ефективності технологій. Порівняно різні технології, проаналізовані архітектури систем розпізнавання, що використовуються у глибокому навчанні, проведено їх порівняльний аналіз. Розглянуто основні етапи отримання та обробки даних для подання в нейрону мережу. Розкрито проблеми повязані з розпізнаванням динамічних жестів, виконано постановку задач дослідження.

1.1 Аналіз предметної області реалізації людино-машинної взаємодії за допомогою жестів

Прогрес в області ЛМВ дозволяє зробити спілкування між людьми і всіма видами пристроїв, що базуються на датчиках, більш природним, інтуїтивно зрозумілим, таким, що нагадує людське спілкування. Для цього використовуються різноманітні методи взаємодії. Серед традиційних засобів, все частіше зустрічаються дослідження і розробки по використанню нових, альтернативних методів. Альтернативні методи повинні надавати більш природне людино-машинне спілкування. Людино-машинне спілкування, близьке до природного, вимагає інструментів і функцій, які імітують принципи людського спілкування. Одним з методів людської взаємодії є використання жестів. Сфера застосування технологій розпізнавання жестів рук дуже широка і за останні 2-3 роки було досягнуто значного прогресу. Для побудови обґрунтованої бази для оцінки поточної ситуації та технологій розпізнавання жестів рук, аналіз проводився у наступних трьох напрямах (рис. 1.1): (1) галузі застосування ЛМВ за допомогою жестів; (2) технології розпізнавання жестів рук,

зокрема методи отримання жестів, методи вилучення ознак та методи класифікації жестів; (3) проблеми і виклики щодо реалізації та впровадження систем, що працюють на основі розпізнавання жестів рук.

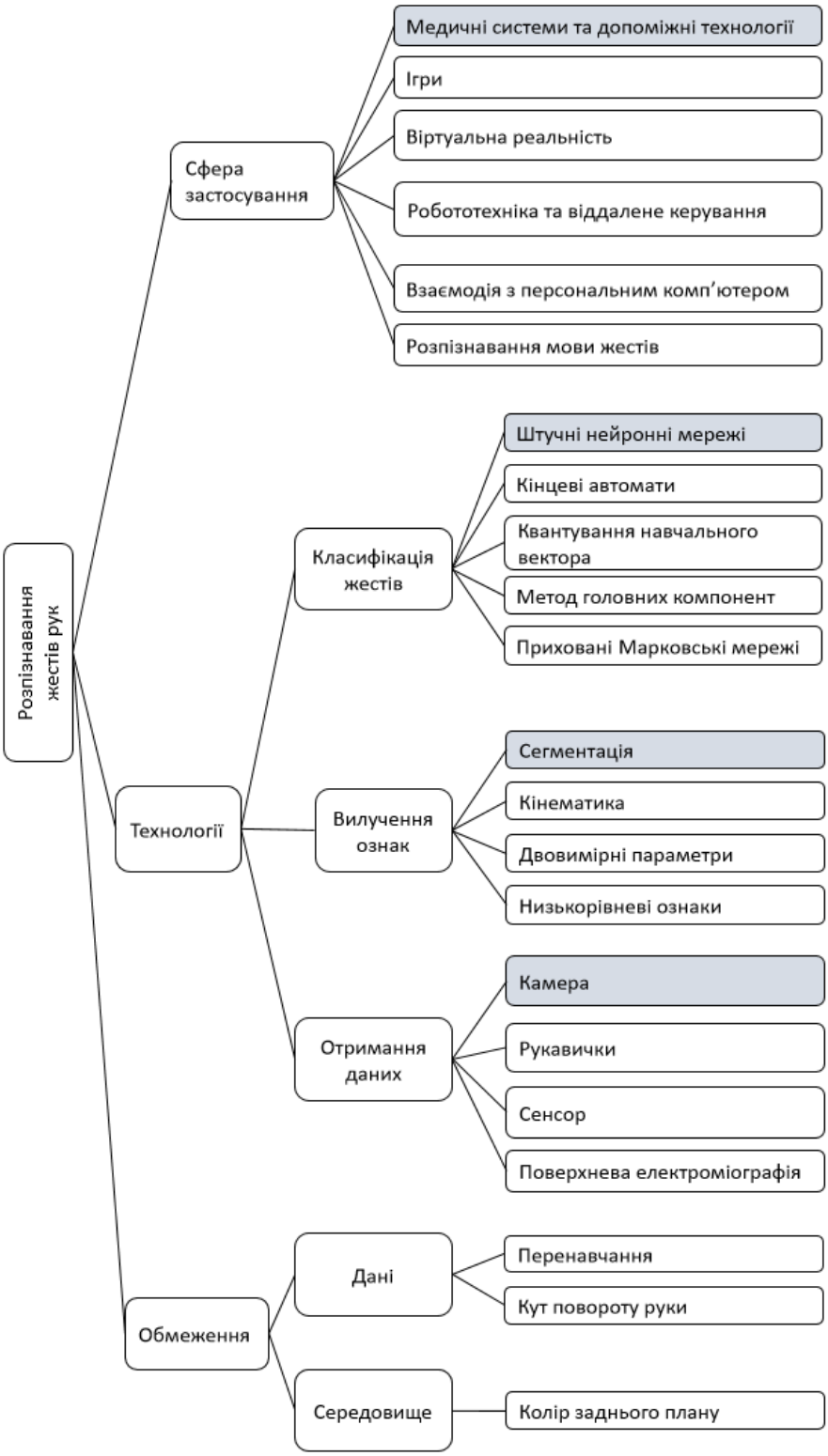


Рисунок 1.1 – Напрями вивчення питання людино-машинної взаємодії з використанням жестів (Адаптовано з [101])

1.1.1 Сфери застосування людино-машинної взаємодії з використанням жестів

1.1.1.1 Загальна таксономія сфер використання технології розпізнавання жестів

Системи з використанням жестів мають кілька галузей застосування, які можна визначити наступним чином [50, 93, 101]:

– Робототехніка та віддалене управління. Віддалене управління технікою за допомогою жестів досить широко використовується у військовій і космічній сферах [30], але має реальні можливості для застосування в багатьох інших областях діяльності - там де людині важко або небезпечно перебувати. Жести використовуються для взаємодії з роботом і управління ним в режимі реального часу [82].

– Для розпізнавання мови жестів. На сьогоднішній день розроблено чимало систем розпізнавання мови жестів для різних мов [18], зокрема української мови [3, 4, 6]. Подібні системи можуть використовуватися не тільки в якості перекладачів але, а й служити в якості в якості невербальних каналів комунікації [2], альтернативним засобом введення тексту в комп'ютер, що для людей з обмеженими можливостями спілкування може бути простіше і природніше, ніж клавіатура [1, 3, 8]. Крім того, з огляду на те, що мови жестів є високо структурованими, вони, в цілому, добре підходять для перевірки того чи іншого візуального алгоритму розпізнавання динамічних образів.

– Для взаємодії з персональним комп'ютером. Деякі дослідники [87] вважають, що жестове управління комп'ютером може надати альтернативу клавіатурі та миші, і активно працюють в цьому напрямку. У [84] управління комп'ютером для вирішення завдань маніпулювання графікою або редагування документів включає в себе використання жестів, що відображають рухи руки.

– Віртуальна реальність. У віртуальній і доповненій реальності використовуються жести, що дозволяють виробляти реалістичні маніпуляції з

віртуальними об'єктами через тривимірний інтерфейс [80] або двовимірний інтерфейс, що імітує тривимірну взаємодію [28].

– Ігри. Жести рук гравця або рухи його тіла можуть бути використані для контролю руху і положення ігрових інтерактивних об'єктів, таких, як автомобіль [25], або аватар [45] у віртуальному світі.

– Медичні системи та допоміжні технології [31, 49], кризове управління та допомога при лихах [76]. Системи людино-машинної взаємодії з використанням жестів рук надають переваги перед звичайними системами машинної взаємодії, і є особливо важливими в області медичних додатків.

Оскільки основна увага даного дисертаційного дослідження приділяється застосуванню жестів в якості допоміжної технології в умовах операційної зали (ОЗ), у наступному підрозділі цей напрям розглянуто більш докладно.

1.1.1.2 Використання жестів в операційних залах

Перспективним сучасним напрямом застосування систем безконтактної ЛМВ є їх використання при наданні медичної допомоги безпосередньо в операційній залі (ОЗ) [69]. Інтраопераційна візуалізація включає:

– використання методів візуалізації в режимі реального часу під час хірургічної процедури;

– аналіз передопераційних знімків;

– залучення супутньої інфраструктури, необхідної для ефективного застосування технологій перегляду.

Центральною проблемою інтраопераційної візуалізації насьогодні є труднощі, з якими стикаються хірурги при отриманні інформації з пристроїв візуалізації в ОЗ. Хірургам потрібні зображення, інтерактивні та тривимірні формати. Вони також повинні мати можливість інтегрувати та маніпулювати цими зображеннями під час хірургічної процедури [71]. В даний час доступний цілий спектр інтраопераційних технологій візуалізації, однак можливості їх інтеграції обмежені. Отже, ці технології візуалізації все ще не знайшли широкого використання. Потрібні вдосконалення вимагають від систем людино-машинної

взаємодії забезпечувати (1) якісний доступ хірургів до інтегрованих зображень, отриманих до та під час хірургічних процедур, (2) навчання персоналу ОЗ, (3) виконання завдань планування та моделювання, (4) ефективну роботу в мультимодальному середовищі.

Технічні пріоритети для вирішення цих клінічних потреб мають зосереджуватися на вдосконаленні самих систем взаємодії та візуалізації шляхом розробки технологій, здатних забезпечити:

1) легкий та зручний доступ до перегляду 2D та 3D-зображень в режимі реального часу в ОЗ;

2) систем, що інтегрують поточні та майбутні системи візуалізації;

3) стандартизацію методів людино-машинної взаємодії в ОЗ.

Умовою застосування систем безконтактної людино-машинної взаємодії (ЛМВ) в ОЗ є підтримка мобільності оперуючого лікаря як в плані пересування по операційній кімнаті, так і в плані збереження тактильної мобільності. Технологією, що виконує умови мобільності пересування та тактильної мобільності, є системи людино-машинної взаємодії з використанням жестів (рис. 1.2).

Рисунок 1.2 – Приклад реалізації ЛМВ в інституті передової біомедичної інженерії та науки Токійського жіночого медичного університету [24]

Системи ЛМВ на підставі відеозапису з використанням жестів використовуються для взаємодії з медичними приладами [104], управління дисплеями візуалізації медичних зображень [27]. Деякі з цих концепцій були використані для поліпшення медичних процедур і систем [67]. Реалізована через подібну систему ЛМВ дозволяє безконтактно керувати комп'ютером, віддаленою машиною або роботом, використовуючи камеру або інший вхідний пристрій для отримання даних про жести рук. Під розпізнаванням жестів рук розуміється інтерпретація комп'ютером зображень або відеопотоку, що подаються на вхід, і присвоєння ймовірності приналежності виявлених об'єктів тому чи іншому класу ЛМВ, що реалізується за допомогою машинного навчання, і дозволяє дистанційно керувати машинною чи роботом, передаючи дані через спеціальне обладнання або камеру. Таким же чином можливо маніпулювання з об'єктами в середовищі віртуальної або доповненої реальності.

Під час проведення оперативних втручань хірург неодноразово може мати необхідність переглянути клінічні дані пацієнта, а саме зображення рентгенологічного, ультразвукового досліджень тощо. Ця необхідність, зазвичай, потребує контакту з обладнанням перегляду зображень, що є причиною втрати стерильності рук хірурга та може призвести до випадкового інфікування операційної порожнини та пацієнта, що трапляється досить часто і є надзвичайно важливим для життя та здоров’я. Спільнота фахівців медичних технологій прагне зменшити кількість таких випадків, розширивши можливості хірурга в операційній. Таким чином, існує необхідність перегляду медичних зображень в ОЗ за умови дотримання стерильності рук хірурга. Одним з варіантів вирішення цієї задачі є впровадження системи безконтактної ЛМВ для маніпулювання зображеннями в операційній. Система безконтактної ЛМВ надасть можливість проводити перегляд зображень без втрати стерильності рук хірурга, ідентифікуючи і розпізнаючи жести для виконання певних команд для маніпуляції зображеннями, такі як, наприклад, збільшення, зменшення розміру перегляду медичного зображення, перегортання зображень тощо.

1.1.2 Технології розпізнавання жестів рук

Для досягнення природної взаємодії людини з комп'ютером, людську руку можна розглядати як пристрій введення. Разом з тим, виразність жестів рук все ще не повністю вивчена для програм ЛМВ [68, 73, 91]. Проблему навчання можна сформулювати наступним чином: враховуючи приклади з набору обмеженого розміру, знайти стислий опис даних [14].

У загальному виді, задача розпізнавання складається з ідентифікації пікселів на зображенні, що становлять руку, вилучення об’єктів з цих ідентифікованих пікселів для класифікації руки та використання цих функцій для розпізнавання появи певних послідовностей пози як жестів.

Підходи, що застосовуються для вирішення проблеми розпізнавання жестів рук підпадають або до технологій, заснованих на традиційному машинному навчанні, або до методів глибокого навчання (рис. 1.3).

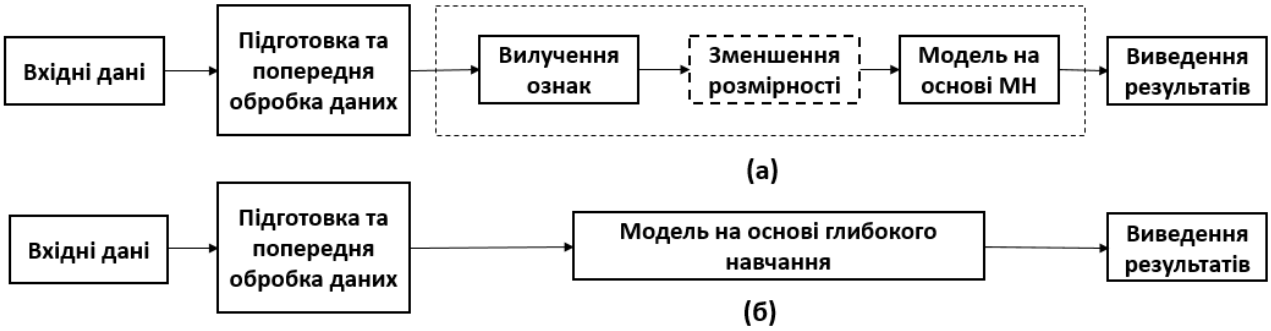


Рисунок 1.3 – Процес розпізнавання жестів (а) на основі традиційного машинного навчання, (б) на основі глибокого навчання (Адаптовано з [17])

Для реалізації машинного навчання необхідно спочатку визначити ознаки, а потім використовувати алгоритми машинного навчання, наприклад, метод опорних векторів (Support Vector Machine, SVM) для класифікації жестів. При використанні методів глибокого навчання, вилучення ознак та моделювання проводиться з використанням нейронних мереж, наприклад, згорткової нейронної мережі (Convolutional Neuron Network, CNN).

З огляду на процеси, означені на рис. 1.3, роботу системи розпізнавання жестів можна розділити на певні етапи, кількість яких може відрізнятися, в залежності від використовуваних алгоритмів, призначення програми та цілей, які ставлять перед собою розробники. Узагальнена схема етапів розпізнавання жестів з візуалізацією стадії підготовки даних представлена на рис. 1.4 [79].

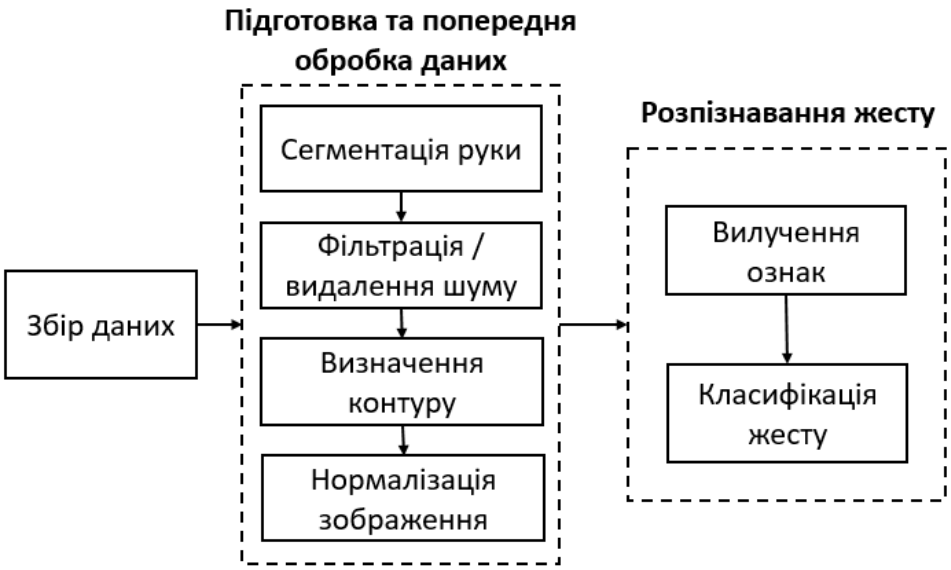


Рисунок 1.4 – Схема технології розпізнавання жестів з акцентом на підготовку даних

Далі етапи збору даних, підготовки та передобробки даних та безпосередньо розпізнавання жестів розглянуто більш докладно.

1.1.2.1 Отримання даних про жести

Підходи, використовувані для збору даних про жести, можна об’єднати у три різні категорії [44]: (1) контактні методи – методи, засновані на використанні рукавичок або зовнішніх датчиків, прикріплених до користувача для захоплення жестів; (2) відстеження з використанням пристроїв з якими людина контактує, наприклад, маніпулятора, миші; (3) візуальні методи та системи, що захоплюють жести за допомогою відеокамери і обробляють їх за допомогою комп'ютерного зору. Остання категорія використовує відеокамеру для захоплення та ідентифікації жесту. Процес розпізнавання використовує

зображення для вилучення серед інших деяких ознак, таких як рух, положення, швидкість, колір.

В контактних методах використовуються спеціально обладнані рукавички або маркерна система. В маркерній системі на руку людини наносяться кольорові мітки, за допомогою яких передаються дані про рух. На основі отриманих даних будується тривимірна модель, що дозволяє з високою точністю відобразити жести. При використанні маркерів в невізуальних підходах застосовується побудова моделі кисті руки. Для опису моделі вказується довжина сегментів і значення кутів між ними. В результаті виходить скелет дискретної фігури, що представляє собою пов'язану множину точок. Моделювання руки з урахуванням всіх її ступенів свободи є складним завданням, тому зазвичай використовується спрощена модель з 27-ю ступенями свободи. Система відстеження положення руки в просторі використовує заздалегідь створену базу даних різних конфігурацій рук. Tomasi і співавторам [90] вдалося створити систему відстеження руки в просторі на основі тривимірної моделі руки, яка може описати форму і зчленовану структуру руки. Цілком очевидно, що, чим більше кількість кутів представленої фігури, тим нижче швидкість її побудови. Вирішується це видаленням малозначних гілок і застосуванням апроксимації, найчастіше через алгоритм Дугласа-Пекера. Іноді різні жести можуть мати схожий скелет, і тоді застосовуються моменти Hu [35] які інваріантні до масштабування і обертання. Рукавички (рис. 1.5) також дозволяють точно з високою швидкістю відобразити природну механіку людської руки і можуть бути оснащені тактильними, вібраційними датчиками, акселерометром та ін.

Так, наприклад, забезпечена сенсорами рукавичка використовувалась в проекті CyberGlove [78], щоб захопити рух і витягти жест засобами апаратного забезпечення рукавички. У [74], слідуючи тому ж підходу, використовували рукавички, приділяючи основну увагу становищу суглобів. Безконтактний інтерфейс Fingual [26] являє собою систему, спрямовану на переклад мови глухонімих в письмову мову.

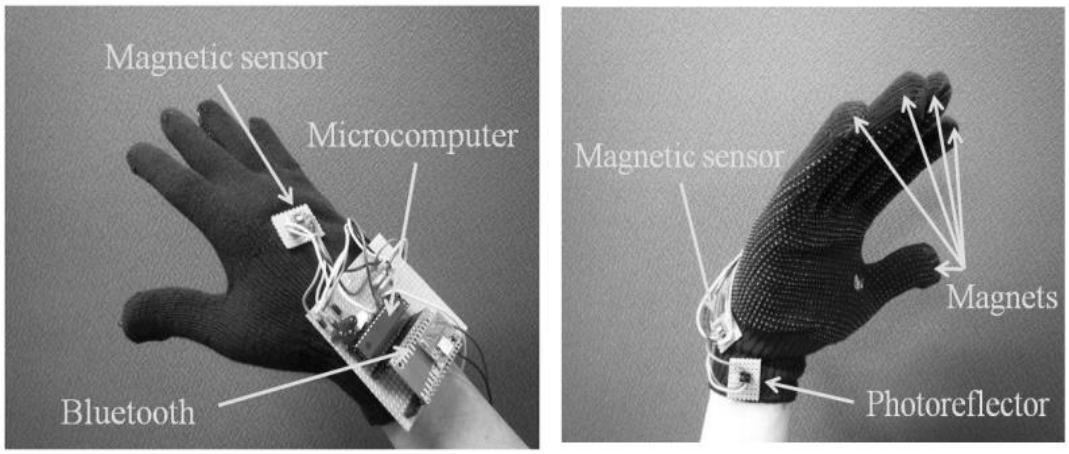


Рисунок 1.5 - Рукавичка Fingual для збору даних про жести [26]

Недоліком даного способу отримання інформації про жести є деяка скутість рухів, оскільки рукавички зазвичай досить громіздкі, і, крім цього, людині властиво використовувати пристрої-посередники для жестового спілкування. Крім того, існує велика залежність від індивідуальної манери висловлювати жести, тому потрібно калібрування під окремо взятого користувача. Виходячи з цього, більш широке застосування отримали візуальні методи представлення жестів. Двома основними підкатегоріями даного підходу є методи, засновані на тривимірній моделі та методи, засновані на зовнішній подібності (рис. 1.6).

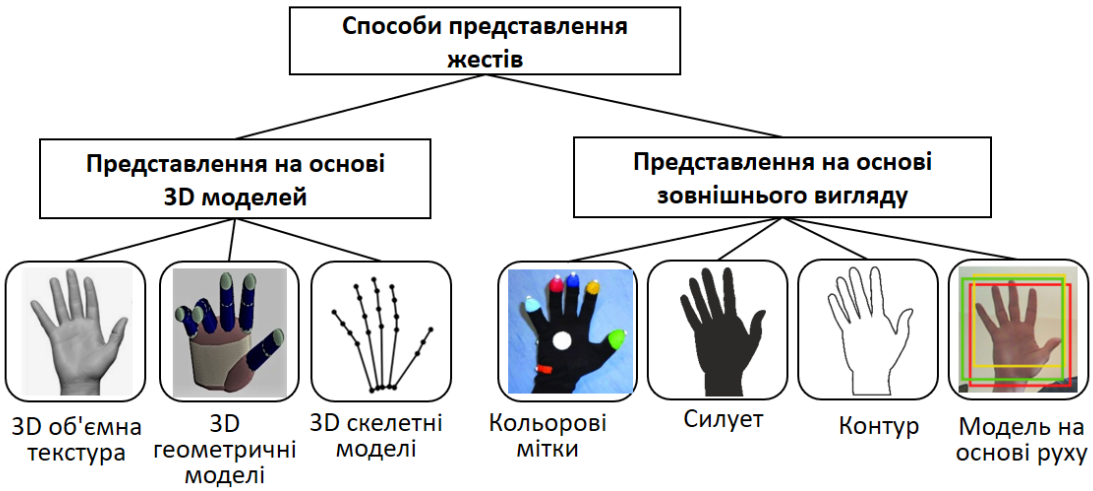


Рисунок 1.6 – Візуальні методи представлення жестів [12]

Дані методи засновані на тому, як людський зір сприймає інформацію про видимі об'єкти. Інформація на комп'ютер, зазвичай, надходить через одну або дві камери - як звичайні, так і інфрачервоні, що вимірюють глибину простору. Характеристиками, що використовуються при ідентифікації жесту, служать контури і силуети. Моделі на основі геометрії силуету можуть включати в себе такі ознаки, як: периметр, опуклість, поверхня, що обмежує прямокутник або еліпс, протяжність, центр ваги і орієнтація.

Моделі, засновані на обрисах, що деформуються, в основному використовують мінливі контури, параметризовані рухом. Моделі на основі руху використовуються для розпізнавання об'єкта на рухомій послідовності зображень. Так, наприклад, у роботі [57] для аналізу просторово-часової активності використовувались гістограми руху.

Головною проблемою візуальних методів розпізнавання жестів є зміна фону і освітлення, при чому знижується якість переднього плану і сегментованого зображення жесту. Для вирішення цієї проблеми в роботі [105] запропоновано використовувати 3σ-принцип нормального розподілу, виходячи з різниці суміжних відеокадрів. Для вибору оптимального порогу пропонується адаптивний метод автоматичного вибору, заснований на методі максимальної дисперсії між класами. Досить великим інтересом серед дослідників [52] користується Microsoft Kinect, обладнаний камерою глибини простору. Завдяки лазерному проектору і сенсорам CMOS він здатний працювати в будь-яких умовах освітлення. Часто використовується поєднання відразу декількох способів отримання даних: через колір, глибину простору, каркас скелета і т.д.

1.1.2.2 Підготовка та попередня обробка даних

Процес вилучення ознак є ключовим для розпізнавання жестів руки. Ознаки включають в себе: положення руки, долоні, пальців, напрямок і кути, під яким розташовані фаланги пальців. Ознаки можуть бути геометричними - контур руки, кінчики пальців, самі пальці, і не геометричними - колір, силует, текстура.

При вилученні ознак може використовуватися кінематичний підхід, заснований на моделі, при якій положення долоні і кути, під якими розташовані суглоби, приймаються в якості ознак [23]. Суть моделі полягає в пошуку кінематичних параметрів, що переводять двовимірну проекцію на тривимірну модель руки у відповідності з обрізаним по краю зображенням руки. Дані надходять з двох або більше камер. Були показані хороші результати при моделі руки з сімома ступенями свободи. В іншому підході, заснованому на зовнішньому вигляді руки [32], використовуються двовимірні параметри зображення, які порівнюються з параметрами, виділеними з вхідного зображення. У той же час жести моделюються як послідовність уявлень. Власний простір (eigenspace), що використовується в даному підході, забезпечує ефективне представлення великого набору точок високої розмірності з використанням малого набору ортогональних базисних векторів. Ці вектори охоплюють підпростір навчального набору, званого власним простором, таким чином, що лінійна комбінація цих зображень може бути використана для реконструкції зображень з навчального набору. Ідея ще одного підходу, заснованого на низькорівневих ознаках, виходить з того, що зовсім необов'язково реконструювати всю руку. Замість цього було запропоновано витягувати з зображення низькорівневі параметри, які стійкі до шумів, і можуть бути швидко вилучені. До таких параметрів можуть належати: центр долоні [64]; осі, що визначають еліптичну область руки [85], і оптичний афінний потік руки [100]. Одні з перших дослідників, які домоглися успіху при даному підході [85], використовували в якості низькорівневих ознак відтінок шкіри, положення руки по x і y, кут осі найменшої інерції і еліпс що обмежує. Комбінуючи набір низькорівневих ознак через мережу HMM, при розпізнаванні американської мови жестів вони домоглися точності в 97%. Виділені ознаки можуть бути збережені на етапі навчання системи у вигляді шаблонів або можуть бути включені в уже наявну нейронну мережу, HMM або дерева рішень, що мають обмежену пам'ять.

Сегментація є одним з перших кроків підготовки даних, що представляє собою процес виділення об'єкта по його кордонах. Технології вилучення ознак об’єктів на відео складаються з двох типів: (1) методи, що використовують для відслідковування переміщень об’єкту граничне поле або обмежувальний прямокутник (bounding box) та (2) технології семантичної сегментації об’єктів (рис. 1.7). У більшості існуючих технологій сегментація є основним кроком для розпізнавання жестів рук, оскільки вона зменшує надлишкову інформацію з фону зображення, перш ніж передавати їх на етапи розпізнавання [20]. При підготовці даних, мета сегментації зображення - позначити кожен піксель зображення відповідним класом. Очікуваний вихід при використанні технології семантичної сегментації - мітки та зображення високої роздільної здатності, в яких кожен піксель віднесений до певного класу. Ця проблема є складнішою, ніж задача виявлення об'єкта, де потрібно передбачити поле навколо об'єкта, але дещо простіша, ніж сегментація екземплярів, де потрібно не тільки передбачити клас кожного пікселя, але й розрізнити кілька екземплярів одного класу.

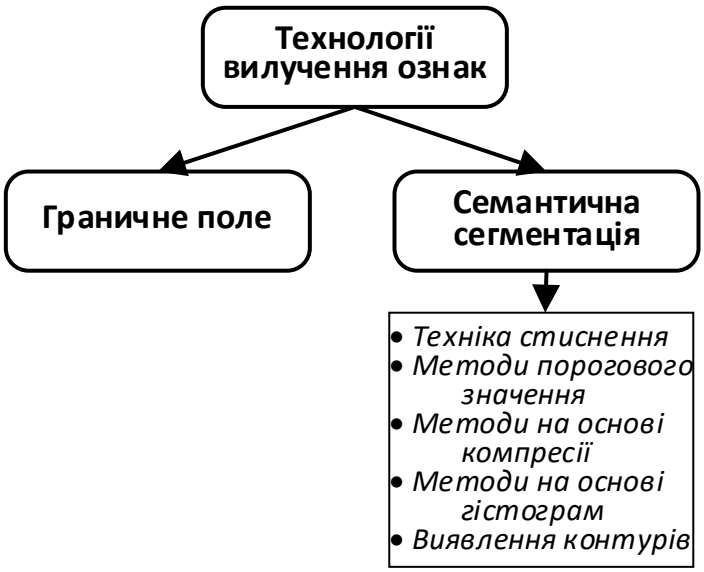


Рисунок 1.7 – Технології вилучення ознак об’єктів на відео

Серед технологій вилучення ознак метод граничного поля є одним з найпростіших, він заснований на розрахунках рівнів порогового значення при перетворенні сірого зображення у бінарне. При застосуванні методу граничного

поля для вилучення ознак, вхідне зображення представлене на сітці розміром S × S. Кожна комірка сітки передбачає наявність відповідного об’єкту. Граничне поле використовується, щоб обмежити об'єкт. Кожне граничне поле складається з 5 елементів: (*x*, *y*, *w*, *h*), де *x* та *y* – координати прямокутника, а *w* та *h* його висота та ширина. П'ятий елемент – це показник достовірності поля який відображає достовірність, ймовірність того, що у граничному полі міститься очікуваний об'єкт та наскільки точні його межі.

Виходячи з особливостей моделі сітки, при використанні граничного поля можна отримати декілька граничних полів на одну комірку сітки. Необхідно щоб лише одна з них відповідала за об'єкт. Для цього вибирається поле, що має найвищий коефіцієнт IoU (intersection over union - перетин над з'єднанням). Кожен прогноз стає кращим при прогнозуванні певних розмірів і співвідношення сторін. Функція втрат складається з трьох частин:

1) Якщо шуканий об'єкт виявлено, розраховується втрата класифікації в кожній комірці як помилка квадрата умовної ймовірності класу відносно до кожного класу.

$$\sum_{i=0}^{S^2} \mathbf{1}_i^{obj} \sum_{c \in classes} (p_i(c) - \hat{p}_i(c))^2 . \qquad (1.1)$$

У цій формулі $\mathbf{1}_i^{obj}$ дорівнює 1 якщо об'єкт є у *i* комірці, інакше 0. $\hat{p}_i(c)$ – умовне позначення вірогідності класу *c* у *i* комірці.

2) Оцінка локалізаційних втрат вимірює помилки у передбачуваних місцях та розмірах вікон.

$$\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbf{1}_{ij}^{obj} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2] +$$

$$+ \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbf{1}_{ij}^{obj} [(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i}\right)^2] \ , \qquad (1.2)$$

де $\mathbf{1}_{ij}^{obj}$ дорівнює 1 якщо *j* граничне поле відповідальне за виявлення *i* об'єкту, інакше 0. λ_{coord} – збільшує вагу втрати для координат граничного поля.

Помилка у 2 пікселі може мати різну вагу залежно від розміру поля. Щоб частково вирішити це, мережа використовує квадратний корінь ширини та висоти граничного поля замість ширини та висоти. Крім того, щоб приділити більшу увагу на точності граничного поля, втрата помножується на додатковий коефіцієнт λ_{coord} .

3) Втрата впевненості розраховується для виявленого об'єкта до якого втрачена довіра, іншими словами вимірює об'єктивність обраного для поля класу.

$$\sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbf{1}_{ij}^{obj} (C_i - \widehat{C}_\iota)^2, \tag{1.3}$$

де C_i – оцінка достовірності граничного поля. $\mathbf{1}_{ij}^{obj}$ як і у випадку з втратою локалізації дорівнює 1 якщо *j* граничне поле відповідальне за виявлення *i* об'єкту, інакше 0. Якщо довіра до об'єкта не була втрачена то втрата розраховується за іншою формулою.

$$\lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbf{1}_{ij}^{noobj} (C_i - \widehat{C}_\iota)^2, \tag{1.4}$$

де $\mathbf{1}_{ij}^{noobj}$ – доповнення $\mathbf{1}_{ij}^{obj}$, C_i – оцінка достовірності граничного поля. λ_{noobj} – коефіцієнт зниження втрати при розпізнаванні фону.

Більшість полів не містить жодних об'єктів. Це викликає проблему дисбалансу класів, тобто модель навчається виявляти фон частіше, ніж виявляти об'єкти. Щоб виправити це, втрата зменшується на коефіцієнт λ_{noobj}.

В результаті об'єднання всіх трьох частин отримаємо повну функцію втрат:

$$\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbf{1}_{ij}^{obj}\left[(x_i-\hat{x_\imath})^2+(y_i-\hat{y_\imath})^2\right]$$
$$+\ \lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbf{1}_{ij}^{obj}\left[\left(\sqrt{w_i}-\sqrt{\hat{w_\imath}}\right)^2+\left(\sqrt{h_i}-\sqrt{\hat{h_\imath}}\right)^2\right] \quad (1.5)$$
$$+\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbf{1}_{ij}^{obj}\left(C_i-\hat{C_\imath}\right)^2+\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\mathbf{1}_{i}^{noobj}\left(C_i-\hat{C_\imath}\right)^2$$
$$+\sum_{i=0}^{S^2}\mathbf{1}_{i}^{obj}\sum_{c\in classes}\left(p_i(c)-\hat{p_\imath}(c)\right)^2$$

Перевагою використання граничного поля для вилучення ознак та розпізнавання жестів є велика швидкість в порівнянні з іншими методами, цей метод добре підходить для використання в реальному часі. Успішність методу полягає у виборі порогових значень. Інші дослідження [5, 22, 92] використовують багатовимірні нечіткі правила які створюють нелінійні пороги. Рішення щодо належності кожного пікселя до сегменту ґрунтується на багатовимірних правилах, що випливають із нечіткої логіки та еволюційних алгоритмів, заснованих на середовищі та освітленні зображення.

Ідея методу сегментації, заснованій на компресії полягає в тому, що оптимальною сегментацією вважається така, що мінімізує за всіма можливими сегментаціями довжину кодування даних. Цей метод намагається знайти шаблони в зображенні, і будь-яка закономірність зображення може бути використана для його стиснення. Метод описує кожен сегмент за його фактурою та формою границь. Кожен з цих компонентів моделюється функцією розподілу ймовірностей, і його довжина кодування обчислюється. Для будь-якої заданої сегментації зображення алгоритм розраховує деяку кількість бітів, необхідних для кодування цього зображення на основі заданої сегментації. Таким чином, серед усіх можливих сегментацій зображення, мета - знайти сегментацію, яка

створює найкоротшу довжину кодування. Цього можна досягти агломераційним методом кластеризації [72].

Методи на основі гістограм є дуже ефективними порівняно з іншими методами сегментації зображень, оскільки вони, як правило, потребують лише одного проходу через пікселі. У цій техніці гістограма обчислюється з усіх пікселів зображення, а піки та долини в гістограмі використовуються для визначення кластерів на зображенні. Колір або інтенсивність можна використовувати як міру. Удосконалення цієї методики полягає в рекурсивному застосуванні методу пошуку гістограми до кластерів зображення, щоб розділити їх на менші кластери. Ця операція повторюється з меншими та меншими кластерами до тих пір, поки не сформується більше кластерів. Одним з недоліків методу пошуку гістограм є складність визначення значимих піків і площини на зображенні. Підходи, засновані на гістограмах, також можуть бути швидко адаптовані до застосування до декількох кадрів, зберігаючи ефективність їх одноразового проходження. Гістограма може бути виконана декількома способами, якщо розглянуто кілька кадрів. Той самий підхід, який застосовується з одним кадром, може бути застосований до декількох, і після об'єднання результатів піки та площини, які раніше було важко визначити, є більш імовірними. Гістограма також може бути застосована на основі пікселя, де отримана інформація використовується для визначення найбільш частого кольору для пікселя.

Виявлений контур - це поле, вилучене в межах обробки зображень. Межі регіонів та краї тісно пов'язані, оскільки на кордонах області часто відбувається різке регулювання інтенсивності. Тому методи виявлення контурів можуть бути використані як основа для альтернативної методики сегментації.

У [9] представлено комплексний підхід до вилучення ознак в якому приймалися до уваги динамічно вибрані точки в контурі руки. Тим самим мінімізувалася модель руки і знижувалася розмірність векторе ознак.

1.1.2.3 Методи розпізнавання жестів

Розпізнавання жестів є заключним етапом, на якому значення жесту має бути ідентифіковано. На цій стадії зазвичай працює класифікатор, який відносить кожен жест, що визначено на зображенні, до певного класу з певною ймовірністю його відповідності цього класу. Більш докладний аналіз класифікаторів, використовуваних для розпізнавання жестів рук надано у п. 1.3.

1.1.3 Загальні проблеми і виклики щодо реалізації розпізнавання жестів рук на основі візуальних даних

З точки зору ЛМВ, на ефективність розпізнавання жестів рук впливає оточуюче середовище (світло, фон, відстань, колір шкіри), положення та напрямок рук.

При застосуванні камери, найпростішим і найбільш природним орієнтиром для отримання інформації про руку є колір шкіри, оскільки він не змінюється при масштабуванні, зрушеннях та поворотах [86].

Колірна модель RGB досить чутлива до змін освітлення, що взагалі є проблемою при даному підході, тому її використання є недоцільним. Навпроти, варто використовувати інші колірні простори, наприклад, HSV або YCbCr, в яких хроматична інформація зберігається в окремих каналах. Застосування даних колірних моделей знижує невизначеність з мінливим при різних умовах кольором шкіри, однак, не вирішує її повністю. У таких випадках можливе використання моделі змішування Гауса (Gaussian Mixture Model) [55], яка використовується при навчанні моделі для стійкого відділення всіх кольорів, що не відповідають кольору шкіри. Слід зазначити, що колір шкіри може відрізнятися у різних рас і етносів, тому даний підхід не можна вважати універсальним.

Іншим недоліком моделей розпізнавання жестів на основі візуальних даних є те, що великі нейронні мережі, навчені на відносно невеликих наборах, можуть перевершувати навчальні дані. Це призводить до того, що модель вивчає

статистичний шум в навчальних даних, наслідком чого є зниження продуктивності, коли модель використовується для прогнозування / генерації значень наступного кадру. Помилка узагальнення збільшується через перенавчання моделі. Вирішення цієї проблеми запропоновано у Розділі 3.1.

1.2 Аналіз особливих випадків розпізнавання та прогнозування жестів рук у відеопослідовності

В деяких випадках, розпізнавання жестів людини бажано виявити якомога раніше [43, 46], щоб зробити ЛМВ більш природною. Разом з тим, більшість досліджень зосереджені виключно на розпізнаванні жестів рук. Метою розпізнавання жестів є віднесення закінченого жесту з відео до відповідного класу, іншими словами, розпізнавання жестів - це завдання відео аналізу за фактом, яке зосереджено на теперішньому стані. Однак, мета розпізнавання полягає не тільки в тому, щоб знайти цільове зображення в реальному часі і відокремити його від фону, а й проаналізувати динамічні просторово-часові характеристики, відстежуючи початок і кінець розпізнаного жесту в потоці кадрів.

Прогнозування жесту має на меті передбачення закінчення жесту або наступного жесту, використовуючи неповні початкові відеодані. Задача прогнозування є не менш актуальною і, незважаючи на перевагу систем розпізнавання, для вирішення задачі прогнозування також розроблена велика кількість алгоритмів. Слід зауважити, що для вирішення задачі прогнозування недостатньо розробити модель яка буде вирішувати проблему розпізнавання об’єктів, та модель яка буде вдало прогнозувати дії, їх слід інтегрувати їх між собою. Інтегрування моделей є метою систем безконтактної людино-машинної взаємодії, що використовуватимуть розробленні моделі разом для вирішення загальної задачі. Структура прогнозування жестів у відео зображеннях представлена наступним чином, як це представлено на рис. 1.8.

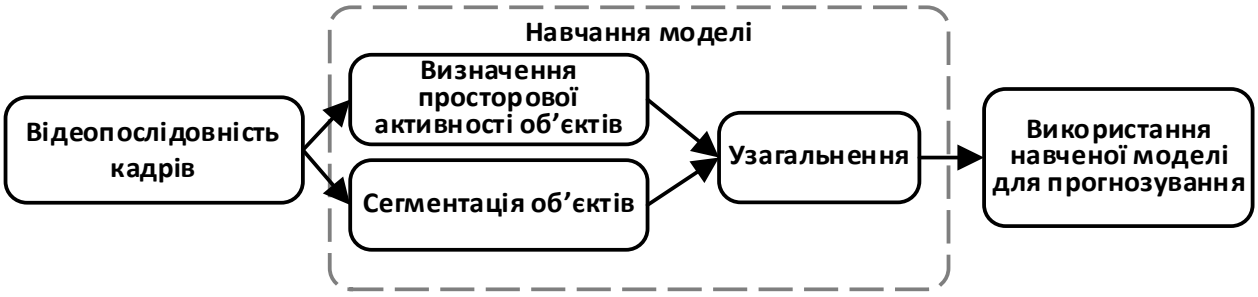


Рисунок 1.8 - Етапи розпізнавання та прогнозування жестів

Підготовка і розпізнавання відеозображень містить наступні етапи.

1. Навчання моделі для розпізнавання об'єктів, що передбачає визначення просторової активності об’єктів з використанням сегментації і даних масок анотованих жестів, цілочисельне значення для кожного зображення та узагальнення – об’єднання узагальненого уявлення жестів.

2. Використання навченої моделі для прогнозування жестів на нових відеозображень в режимі реального часу.

Таким чином, у розпізнаванні та прогнозуванні жестів можна виділити наступні три напрями: (1) розпізнавання статичних жестів, (2) розпізнавання динамічних жестів, (3) розпізнавання та прогнозування відеопослідовностей зображень жестів.

1. Розпізнавання статичних жестів. При статичних жестах рука займає в просторі будь-яке одне конкретне становище, не змінюючи його протягом деякого часу.

2. Розпізнавання динамічних жестів. Динамічні жести можна розглядати як послідовність взаємопов'язаних положень руки. При розпізнаванні динамічних жестів слід брати до уваги не тільки просторові, але і тимчасові ознаки жесту.

3. Прогнозування жестів. Прогнозування проводиться на основі деякої кількості кадрів з відеопослідовності з передбаченням подальшого руху руки.

Основною ідеєю при розпізнаванні жестів є те, що кожний жест має певні ознаки, за якими може бути віднесений до певного класу або групи класів, що допомагають класифікувати жест. При цьому, більш складні системи використовують дуже велику кількість ознак, щоб якомога швидше та точніше

обрати клас для жесту. Що стосується розпізнавання жестів на відео, то додається ще одна важлива ознака – рух.

Завдяки руху на відео можна розпізнавати не тільки жести, але й події. Рух сам по собі є потужною візуальною підказкою. Суть багатьох дій саме в динаміці, та іноді достатньо відстежити рух окремих точок, щоб розпізнати якусь подію.

Рух можна характеризувати як точки сцени, що рухаються відносно камери. З позиції передобробки даних, відеопотік може бути представлений як набір сцен, де кожна сцена має набір об’єктів які слід проаналізувати та класифікувати.

Аналіз відеопотоку має декілька етапів (рис. 1.9), що відрізняються від стандартної схеми розпізнавання (рис. 1.3) рівнями деталізації та наявністю додаткових процедур на етапі 4.

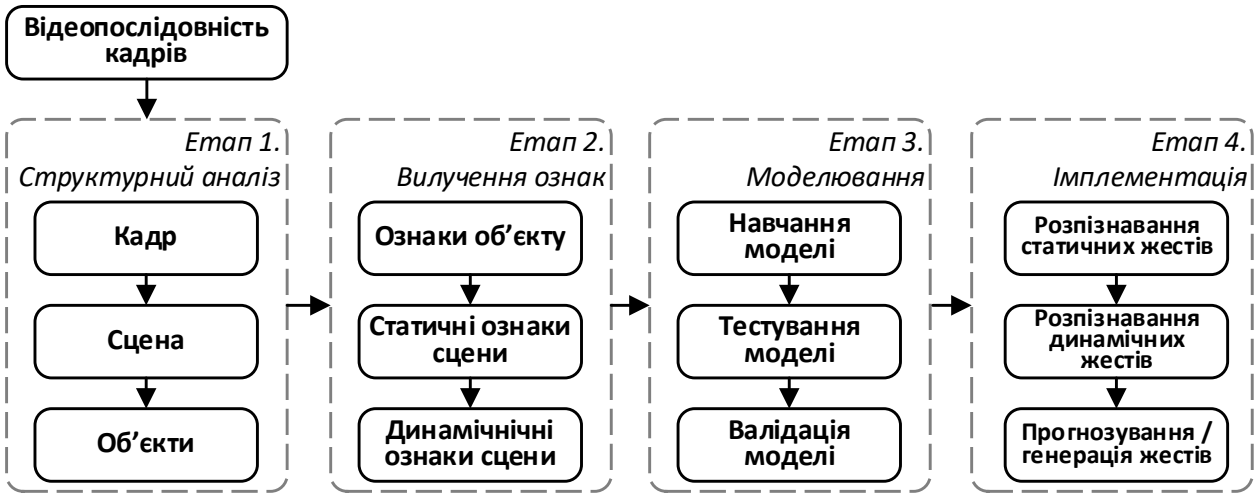


Рисунок 1.9 – Узагальнена схема аналізу відеопотока

Етап 1. Попередня обробка. На цьому етапі вхідний потік відеопослідовності розділяється на кадри, кадри на сцени, а у сцен виділяються об’єкти.

Етап 2. Вилучення ознак. На етапі вилучення ознак проводиться виділення інформації зі структурних одиниць відеопотоку: статичні ознаки сцени, ознаки, які характеризують жест, та ознаки пов’язані з рухом жесту.

Етап 3. Моделювання. Вилучені ознаки використовується для навчання моделі розпізнавання жестів, її тестування та валідації.

Етап 4. Імплементація. На цьому етапі навчені моделі можуть використовуватися для розпізнавання статичних, динамічних жестів та на підставі початкових кадрів прогнозувати/генерувати динамічне закінчення розпізнаного жесту.

1.3 Аналіз архітектур систем розпізнавання жестів

Залежно від підходів та цілей завдання розпізнавання жестів можуть бути вирішені різними методами. Оскільки найбільш ефективними, з точки зору точності розпізнавання візуальних образів і зокрема жестів є глибокі нейронні мережі, у цьому підрозділі основна увага приділена саме їх архітектурам та різним модифікаціям.

1.3.1 Загальний аналіз архітектур глибокого навчання

Методи глибокого навчання можна згрупувати у три категорії: навчання під наглядом, навчання без нагляду та гібридне навчання (рис. 1.10).

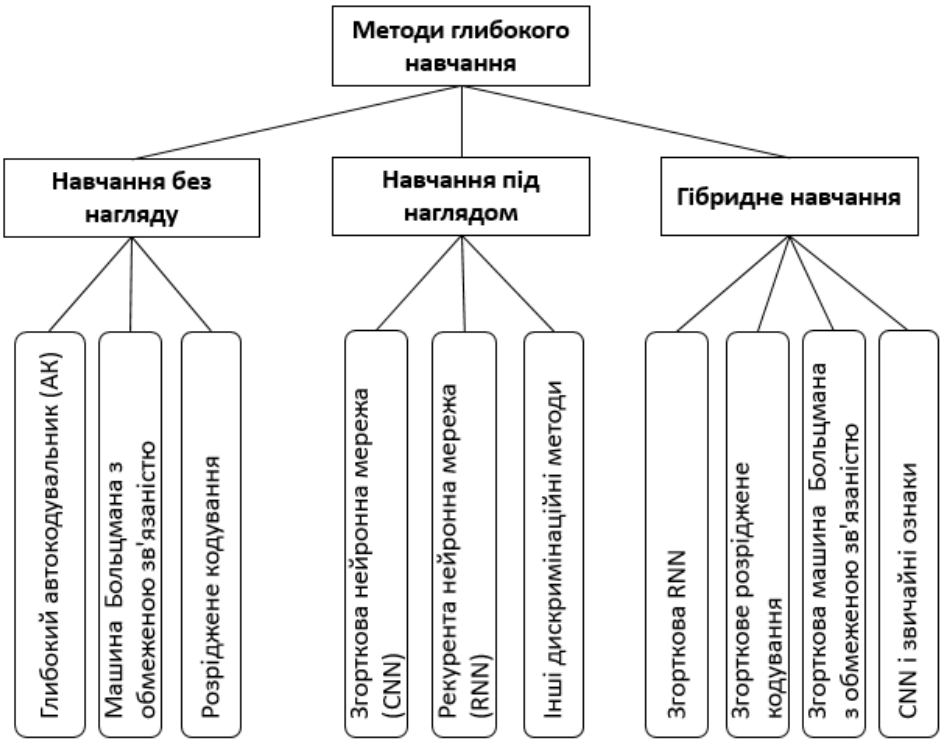


Рисунок 1.10 – Узагальнена таксономія методів глибокого навчання

Перспективні підходи к прогнозуванню, що ґрунтуються на навчанні без нагляду включають різновиди машини Больцмана, розріджене кодування, глибокі автокодувальники (АК), зокрема ймовірнісні моделі АК, такі як Variational Auto-Encoder (VAE) та інші. VAE можуть справлятися з невизначеністю і моделювати безліч можливих майбутніх результатів, однак мають тенденцію давати розмиті прогнози [15, 47, 48]. Іншим прикладом використання навчання без нагляду є самоорганізовані карти Кохонена, що було успішно застосовано у [8] для класифікації наборів даних і перетворення жестів рук у філіппінські слова.

Серед інструментів навчання під наглядом одним з найпопулярніших та найефективніших методів розпізнавання статичних жестів є згорткові нейронні мережі (Convolutional Neural Networks, CNN). CNN має низку переваг перед іншими архітектурами. Перш за все, шари згортки в такій мережі здатні будувати ієрархічні уявлення і самостійно виділяти ознаки у вхідних даних. Однак, для реалізації розпізнавання динамічних жестів в реальному часі, зазвичай більш ефективними є гибридні методи та методи, засновані на інших архітектурах.

Автори [52] використовували CNN для вилучення ознак і SVM для класифікації. Отримана точність склала 98.52%. У [51] CNN з класифікатором softmax на даних RGB-D показала точність класифікації 98.5%.

В [103] для вилучення ознак використовується CNN, а для сегментації зображень класифікатор на основі випадкового лісу (Random Forest). У роботі [46] запропоновані 3D CNN для виявлення жестів в режимі реального часу. Поєднання CNN з рекуррентною нейронною мережею (Reccurent Neural Network, RNN) дозволяє обробляти послідовність зображень або відеопотік, об'єднуючи в одній мережі дані по простору і часу. Значний успіх було досягнуто при використанні в CNN та Microsoft Kinect для розпізнавання італійської мови жестів [75]. П'ять різних архітектур мереж глибокого навчання було протестовано, в результаті чого зроблено висновок, що двонаправлена рекуррентність та часова згортка можуть значно покращити покадрову класифікацію жестів. Ще в 1992 році Yamato та ін. [99] запропонували

використовувати приховану марковську модель (Hidden Markov Model, HMM) для розпізнавання дій людини у відео. З того часу ця ідея неодноразово розвивалася багатьма дослідниками. Зокрема, в [42] HMM використовувався в поєднанні глибоким навчанням для розпізнавання часових жестів. Автори [66] у своїй роботі об'єднали 3D CNN та багатопотокову довгу короткочасну пам'ять (LSTM) LSTM-RNN для визначення та класифікації жестів. Завдяки цій комбінації стало легше обробляти жести змінної довжини. Дослідження в області розпізнавання окремих одиниць жестів рук за останні шість років представлені у табл. 1.1.

Таблиця 1.1 – Результати розпізнавання окремих одиниць жестів рук

Рік	Дослідження	Алгоритм	Результат класифікації
2014	Pigou [75]	NN	Точність 91.7% Індекс Жаккарда of 0.789
2014	Simonyan[83]	SVM	Точність 88%
2015	Molchanov[58]	CNN	Точність 94.1%
2016	Nishida [66]	LSTM	Точність 97.8%
2017	John[41]	CNN	Точність 90%
2017	Bheda[10]	CNN	Точність 82.5-97%
2017	Hussain[36]	CNN	Точність 93.09%
2018	Alashhab[7]	CNN	Точність розташування 96%–100%, точність розпізнавання 99.45%
2018	Devineau[22]	CNN, MLP	Точність 91.28%
2019	Köpüklü [46]	ResNeXt-101	Точність в автономному режимі 94,04% та 83,82% для модальності глибини відповідно до тестів EgoGesture та NVIDIA

Автори [10] реалізували метод класифікації зображень як букв, так і цифр для американської мови жестів із приблизно 82,5% точності жестів алфавіту та 97% точності на валідаційній виборці, що складалася з цифр. У дослідженні [7] протестували сім найсучасніших архітектур CNN та представили метод виявлення та локалізації жестів рук, що заснований на глибокому навчанні та

отримав високу точність як у локалізації (96% –100%), так і в розпізнаванні (99,45%). З табл.1.1 видно, що існуючі методи розпізнавання окремих одиниць жестів досягли високої точності, але, в той же час, не до кінця вирішеною задачею залишається розпізнавання динамічних жестів. Справа в тому, що системи реального часу для розпізнавання жестів рук потрібно застосовувати для виявлення та класифікації одночасно на безперервному потоці відео.

Є кілька робіт, що стосуються реалізації завдань одночасного виявлення та класифікації. У роботі [70] пропонується алгоритм на основі гістограми орієнтованого градієнта (Histogram of Oriented Gradient, HOG) разом із SVM-класифікатором. Автори [59] використовують спеціальну радіолокаційну систему для виявлення та сегментації жестів. У роботі [60] використана функція втрат тимчасова класифікація без зв'язку (Connectionless Temporal Classification, CTC) для виявлення послідовних жестів. Однак CTC не забезпечує одноразових активацій.

Таким чином, можна констатувати, що для виявлення просторових ознак статичних жестів досить використання CNN. Що стосується виявлення кореляцій в часі, то краще за все з цим можуть впоратися мережі з довгою короткостроковою пам'яттю. Одним з варіантів реалізації системи з розпізнаванням динамічних жестів є архітектура CNN + LSTM, в якій вектор просторових ознак подається зі згорткової частини мережі на шар LSTM для знаходження в них часових закономірностей. Така архітектура показала високий ступінь точності, але все ж при цьому не можна бути впевненим у тому, що кореляція між просторовими і часовими шарами визначена вірно.

Отже, незважаючи на те, що популярність розпізнавання дозволила досягти високої точності розпізнавання жестів [40, 53, 54], дослідження в галузі прогнозування і генерації жестів менш поширені.

Серед робіт, присвячених прогнозуванню жестів найбільш використовуваними є гібридні архітектури на основі LSTM. Так, наприклад, для задачі прогнозування відеопослідовностей, в [81] запропоновано архітектуру ConvLSTM, в якій блоки LSTM зчитують вхідні дані, використовуючи операцію

згортки на кожному часовому кроці. Як і у випадку комбінації CNN + LSTM, вхідні дані розбиваються на послідовності з фіксованою кількістю часових кроків. Таким чином, виявлення закономірностей в часовій залежності між кадрами відеопослідовності стає більш надійним. Така архітектура може бути використана у прогнозуванні наступних кадрів при розпізнаванні жесту до того, як він буде завершений. Однак, слід зазначити, що при цьому значно збільшується час навчання, яке до того ж потребує значних комп'ютерних ресурсів. У [94] використовувався радіочастотний сенсор для отримання ознак динамічних жестів, які потім передавалися на навчання мережі архітектури CNN і LSTM. Автори [19] розробили генеративну модель, яка вирішує проблему синтезу відеопози і жестів. Ключове нововведення полягає в запропонованій архітектурі локалізованого перетворення, яка фіксує часові та просторові перетворення в областях, що становлять інтерес, і надає інструкції про те, що і де прогнозувати для цих ключових областей в майбутніх кадрах.

Результати проведеного аналізу та отримана точність представлені в таблиці 1.2.

Таблиця 1.2 – Точність моделей прогнозування жестів

Рік	Дослідження	Архітектура мережі	Результат
2016	Lotter [56]	PredNet	MSE=0.00313 SSIM=0.884
2017	Barros [9]	Convexity Approach, HMM and Dynamic Time Warping	Рівень прогнозування = 97%
2017	Xiong [98]	Двонаправлена ConvLSTM	MAE=1.13-2.10 MSE=1.43-7.6
2018	Tang [89]	MSST-ConvLSTM	MAE = 8.2
2019	Wei [96]	CrevNet	SSIM=0.949 MSE=22.3
2019	Cui [19]	adopted U-Net	Точність 62.7-98.5%
2020	Wu [97]	SalSAC	AUC-Judd = 0.898 Міра схожості = 0.480 LCC = 0.729

З аналізу останніх досліджень, де отримано найкращі результати прогнозування (табл. 1.2), було виявлено декілька важливих особливостей, а саме:

(1) Здатність мереж класифікувати жести лише на основі постави руки. Ця можливість дозволяє моделювати будь-який жест, лише будуючи послідовність жестів з різними положеннями рук.

(2) Архітектура мережі дозволяє мати справу з дисперсіями швидкості, і навіть з половиною наявних кадрів може мати коефіцієнт прогнозування вище 60%. Це означає, що модель здатна справлятися з дисперсією швидкості, але не з затримкою часу. Використання рекурентних нейронних мереж, архітектурні особливості яких дозволяють впоратися з тимчасовим запізненням, а саме моделей з довгою короткочасною пам'яттю, потенційно, може допомогти вирішити цю проблему.

(3) Жести можна виконувати з різною швидкістю або з різним положенням рук, але зміни, які змусять модель працювати в реальному часі ще необхідно вивчати додатково [9, 19]. Необхідне, також, розширення системи розпізнавання для інтеграції передбачення та розпізнавання в одній системі.

Таким чином, проведений огляд існуючих технологій показав необхідність створення власних моделей розпізнавання та методу прогнозування жестів хірурга в операційній кімнаті в реальному часі для безконтактного управління переглядом медичних зображень.

Оскільки крім виявлення просторових ознак, завдання розпізнавання та прогнозування жестів рук передбачає також визначення часової послідовності, в якості базової архітектури, має сенс розглянути два типи мереж – стандартну CNN і рекурентну нейронну мережу, зокрема її різновид LSTM.

1.3.2 Особливості архітектури мережі LSTM

LSTM була запропонована у 1997 р. [33] для вирішення проблеми довгострокової залежності. Кожен блок LSTM містить три типи вентилів:

вхідний, вихідний затвор і забувальний, які реалізують функції запису, читання та скидання в пам'ять блоку (рис.1.11).

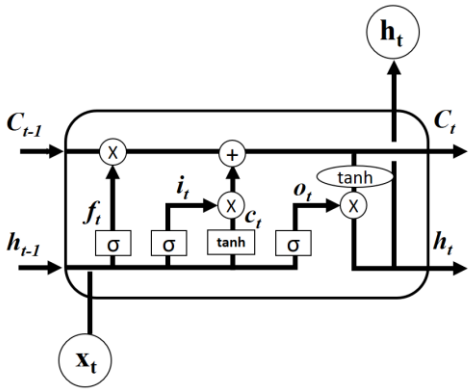


Рисунок 1.11 – Схема блоку LSTM [33]

Поточний вхідний вектор x_t і попередній короткочасний стан h_{t-1} надходять на чотири окремі повнозв'язні шари. Перш за все вирішується: яка інформація не потрібна для подальшої роботи, тобто видаляється частина довгострокової пам'яті. Це рішення приймається на рівні контролю забувального вентиля (forget gate) f_t. В залежності від одиниці чи нуля, що отримується через сигмоїдальну функцію, це означає «повністю зберегти» або «повністю позбутися». Вектор забувального вектору виглядає як:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \tag{1.6}$$

де W_f – рекурентні ваги для вентиля забування, b_f – упередження.

На рівні шару c_t з функцією, що реалізує гіперболічний тангенс, виконується аналіз поточного входу x_t та попереднього короткострокового стану h_{t-1}. На виході маємо вектор значень c_t, котрі можливо будуть додані в довгострокову пам'ять:

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c). \tag{1.7}$$

Які саме значення будуть передані во вхідний вентиль, залежить від результату поелементного множення векторів c_t та i_t. Вектор i_t є виходом із шару з сигмоїдальною функцією та визначається по формулі:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i . \quad (1.8)$$

Після чого, розраховується довгостроковий стан:

$$C_t = f_t \times C_{t-1} + i_t \times c_t . \quad (1.9)$$

Вихідний вентиль обирає, які з частин довгострокового стану слід читати та виводити на цьому етапі часу y_t. Шар вихідного вентиля з сигмоїдальною функцією o_t обчислюється за ф. (1.10):

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) . \quad (1.10)$$

Нарешті, короткочасний стан обчислюється як:

$$h_t = o_t \times \tanh(Ct) . \quad (1.11)$$

Завдяки своїй складній структурі, LSTM здатна вивчати довгострокові залежності оскільки враховує просторово-часові дані і дозволяє проводити вилучення ознак, розпізнавання та прогнозування жестів на підставі відео для реалізації безконтактної людино-машинної взаємодії але, як зазначається в [105] більшість відомих моделей, де поєднуються LSTM та CNN не є моделями в режимі реального часу.

Так, наприклад, у [63] досягнуто дуже високу точність для шести жестів, але кожен жест повторювався до 200 разів, що є надзвичайно великим робочим навантаженням. Крім того, використання розсувного вікна на 940 мс суттєво знижує продуктивність моделі в режимі реального часу.

Частково, вирішення цього питання може бути досягнуто за рахунок використання моделей на основі скінченних автоматів.

1.4 Застосування скінченних автоматів для реалізації людино-машинної взаємодії з використанням жестів

З точки зору реалізації ЛМВ, жести являють собою виразні, значущі рухи тіла що передають інформацію для взаємодії з комп'ютерною системою. Жести рук виконують три функціональні ролі: семіотичну, ерготичну та епістемічну [29]. Семіотична функція полягає в передачі інформації, ерготична функція відповідає здатності маніпулювати об'єктами в реальному світі, а епістемічна функція дозволяє вчитися у навколишнього середовища за допомогою тактильного досвіду. Згідно [77], вся процедура відстеження жестів від їх подання до перетворення у будь-яку цілеспрямовану команду відома як розпізнавання жестів. Відповідно до різних сценаріїв застосування в системах ЛМВ, жести рук можна класифікувати на два типи – статичні та динамічні. Під статичним розуміються жести, що знаходяться в одній конкретній позиції, які не змінюються в часі. Динамічні жести змінюються протягом часу і є послідовністю деяких мінливих позицій. Статичні жести, зазвичай, описуються з точки зору форм рук, а динамічні жести - відповідно до рухів рук [16]. До переваг розпізнавання статичних жестів належать висока чутливість, надійність та простота обчислень. Для порівняння, динамічні жести здебільшого вимагають довшого часу для алгоритмів розпізнавання, оскільки потребують спостереження за рухом жесту від його ініціації до завершення. Виходячи з природи, статичний жест може бути виявлений майже миттєво, одразу після його ідентифікації, але існує можливість помилки через схожість деяких з них, що може призвести до помилкових спрацьовувань.

В залежності від типу, системи ЛМВ за допомогою жестів рук на основі зору, можуть містити різні етапи, включаючи статичні та динамічні форми жестів рук, прогнозування та управління в реальному часі (рис. 1.12). Як

зазначалося вище, найбільш складною проблемою динамічного розпізнавання жестів є його просторово-часова мінливість, коли один і той же жест може відрізнятися за швидкістю, формою, тривалістю та цілісністю [95].

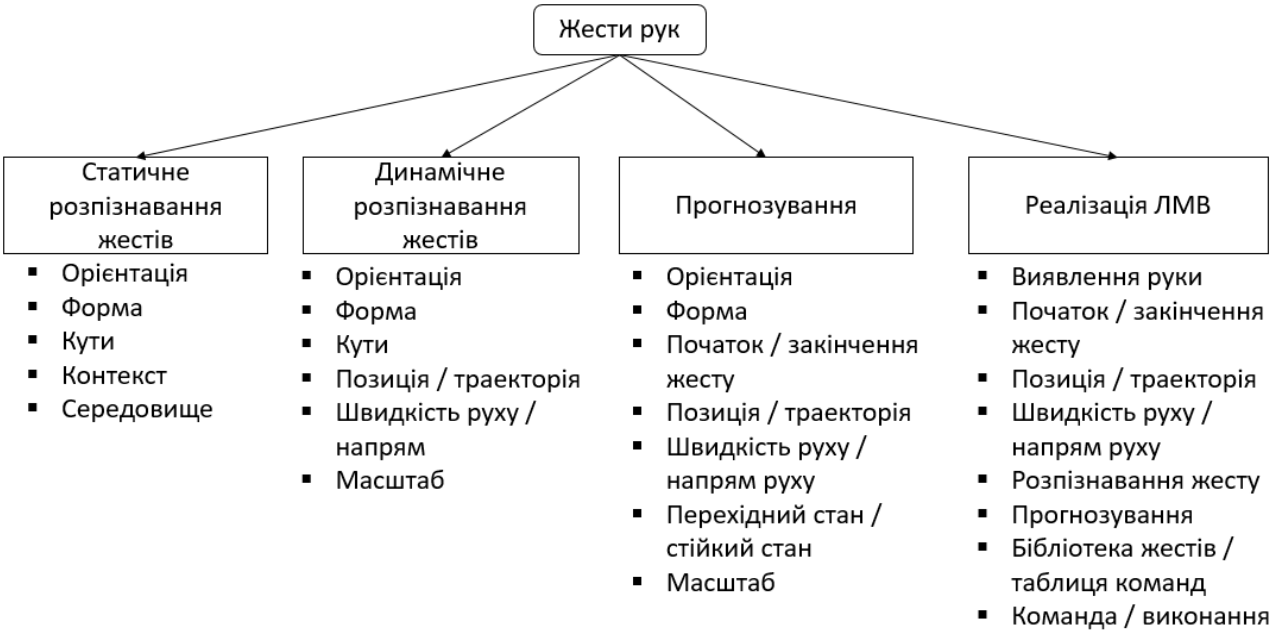


Рисунок 1.12 – Способи використання жестів для реалізації ЛМВ на основі зору та їх основні характеристики

Ці характеристики ускладнюють розпізнавання динамічних жестів рук, але в контексті організації ЛМВ існує ще декілька складностей, наприклад, умова, згідно якої жести повинні розпізнаватися лише один раз (активація в режимі одиночної дії) [46]. Це дуже критично саме для ЛМВ, і ця проблема все ще повністю не вирішена.

У загальному виді, процес розпізнавання жестів у системах ЛМВ можна розбити на такі компоненти (рис. 1.13).

В реалізації ЛМВ на основі зору, жести можна розглядати як кінцеву послідовність станів. Згідно з дослідженням [34], кожен жест може бути визначено як упорядковану послідовність станів з використанням просторової кластеризації і часового вирівнювання. Типовим підходом до опису руху руки є використання моделей скінченного автомата, наприклад, прихованої моделі Маркова (Hidden Markov Model, HMM), для перетворення ряду рухів в опис

команд. HMM - це статистична модель, яка широко використовується для розпізнавання жестів рук [13, 61, 62], оскільки здатна зберегти просторово-часову ідентичність жесту руки і має можливість виконувати автоматичну сегментацію.

Рисунок 1.13 – Процес розпізнавання жестів в системі ЛМВ на основі зору (Адаптовано з [29, 88])

Такі системи працюють, навчаючи HMM аналізувати потік відстежуваних короткострокових рухів. В [102] представлена модель скінченного автомата, яка має п'ять станів, кожному жесту відповідає команда запуску (S), вгору (U), вниз (D), вліво (L) і вправо (L). Результати розпізнавання служать вхідними даними для скінченного автомату для генерації машинних команд, щоб імітувати передбачувану операцію відповідного жесту. У [21] скінченний автомат використовувався для моделювання чотирьох якісно різних фаз жесту руки: (1) статичне початкове положення, (2) плавний рух руки і пальців під час жесту, (3) статичне кінцеве положення, і (4) плавний рух руки назад у вихідне положення. Усі жести були представлені у вигляді списку векторів, які зіставлялися зі збереженими векторними моделями жестів за допомогою пошуку в бібліотеці на основі векторних зсувів.

Характеристики руху кожної часової точки моделювались у більшості методів динамічного розпізнавання жестів руками за допомогою HMM, тим не менше, варто відзначити, що всі символи траєкторії жесту не враховуються одночасно. Крім того, розпізнавання на основі місцевих особливостей дуже чутливе до періоду дискретизації та швидкості, а постійний локальний процес жестів може спричинити помилкове розпізнавання.

Проблема використання скінченного автомату в ЛМВ для управління переглядом зображень полягає в тому, що розпізнавання жестів є складним процесом, оскільки дані, взяті з траєкторії будь-якого даного жесту, є дуже варіативними. Причинами варіативності даних є відмінна частота дискретизації, помилки розпізнавання або шум, і, в першу, чергу, людські відмінності у виконанні жесту, як в часі, так і в просторі. Точність розпізнавання є вагомим фактором, що впливає на якість ЛМВ на підставі відеозаписів жестів. Невирішеною проблемою є прискорення швидкості управління переглядом зображень на підставі прогнозованих кадрів відеопослідовностей жестів за умов попередження помилкових спрацьовувань скінченного автомату. Вирішення цієї проблеми запропоновано і розглянуто у Розділі 3.2.

1.5 Аналіз вимог до систем людино-машинної взаємодії із застосуванням жестів

1.5.1 Аналіз нормативної бази

Основним документом, що регламентує розробку та використання систем ЛМВ з використанням жестів є сімейство стандартів ISO / IEC 30113-1, зокрема ISO / IEC 30113-11: 2017 Інформаційні технології - Інтерфейси на основі жестів між пристроями та методами - Частина 11: одноточкові жести для загальних системних дій [37], ISO / IEC 30113-12: 2019 Інформаційні технології - Інтерфейси користувача - Інтерфейси на основі жестів між пристроями та методами - Частина 12: Багатоточкові жести для загальних системних дій [38].

Стандарт ISO / IEC 30113-1 складається з керівних принципів для проектування інтерфейсів на основі жестів. Стандарт визначає низку вимог та рекомендацій, серед яких: активація / закінчення жесту, виконання жесту, зворотний зв'язок для підтвердження жесту, переадресація, скасування жесту, критерії розміру жесту, контроль критеріїв, зміна відповідності команди жесту та описи окремих жестів у межах виконуваної діяльності.

Один з останніх, прийнятих в цій групі стандартів, ISO / IEC 30113-60: 2020 - Інформаційні технології - Інтерфейси на основі жестів між пристроями та методами - Частина 60: Загальні вказівки щодо жестів для зчитувачів з екрана [39] надає загальні вказівки щодо жестів для зчитувачів з екрана, що працюють на різних ІКТ-пристроях. Він зосереджений на описі жестів та функцій для зчитувачів з екрана, що працюють на пристроях ІКТ. Разом з тим, як зазначається в описі, стандарт не визначає і не вимагає конкретних технологій розпізнавання жестів.

1.5.2 Технологічні та природні обмеження технологій розпізнавання жестів для використання в системах ЛМВ

Незважаючи на те, що жест є природним способом спілкування, при використанні жестів для ЛМВ необхідно враховувати існуючі обмеження технології та людини. Так, наприклад, управління жестами за рухом робота має бути безперервним протягом усього періоду контролю, в якому користувач повинен мати можливість негайно змінити шлях. Рухи робота повинні бути чутливими настільки, щоб користувач міг отримати миттєвий зворотний зв'язок, щоб бути впевненим, що робот працює відповідно до вхідних даних, інакше будь-яка затримка відповіді буде заважати синхронізації рухів.

Це означає, що використання динамічного розпізнавання жестів в таких системах заборонено, оскільки на виконання динамічного жесту зазвичай потрібно понад півсекунди, що стоїть на заваді реакція системи [88].

Що стосується жестів для виконання дистанційного зчитування з екрана, навпаки, динамічний жест дозволяє досягти необхідного рівня взаємодії з мінімальною затримкою, неможливого при застосуванні лише статичних жестів, наприклад, плавна зміна розмірів об’єкта на екрані.

Попри значні дослідницькі досягнення в галузі технологічного розвитку управління жестами, у багатьох процесах розробки все ще бракує врахування ергономічних вимог, що може мати наслідки для здоров'я та досвіду користувачів при тривалому використанні. Ряд досліджень продемонстрували, що жести, якщо вони не розроблені з урахуванням людського фактору, можуть спричинити втому м’язів після тривалого використання [29]. Наприклад, [11] надано аналіз систем ЛМВ за останні 30 років, де підтверджено, що контрольовані жестами користувацькі інтерфейси здатні надавати реалістичні та доступні можливості, які можуть бути доречні для людей похилого віку та людей з обмеженими можливостями, але в деяких випадках результати тестування показали, що користувачі мали помірний та сильний рівень втоми в плечах та передпліччях при тривалому використанні системи. Це доводить важливість створення не лише інтуїтивно зрозумілого набору жестів, але також необхідність враховувати біомеханічні обмеження людини [65]. Використання систем, побудованих виключно на мові жестів ускладнює виявлення точної динаміки виконання, але проблему можна подолати, додавши звичайні елементи інтерфейсу. Крім того, система повинна бути відрегульована таким чином, щоб уникнути помилкових реакцій на рухи, які не були призначені для введення системи.

1.5.3 Аналіз вимог використання жестів для ЛМВ в операційній залі

Хірургічні монітори та дисплеї є невід’ємною частиною середовища гібридної операційної. Ці монітори відображають важливу інформацію, таку як життєві показники пацієнта, хірургічні зображення, рентгенівські знімки, знімки з магнітно-резонансного томографа (МРТ) та ін. Потреба в стерильних

маніпуляціях із зображеннями в ОЗ призвела до розвитку безконтактних систем ЛМВ на основі використання жестів як альтернативного способу заміни периферійних пристроїв, таких як клавіатура, миша та сенсорні екрани, традиційно використовуваних для навігації та маніпулювання послідовним набором зображень МРТ за межами операційної.

Отже, інтерфейс управління в ОЗ передбачає виключно жести рук, оскільки хірург під час операції значно обмежений у виборі дій та рухів. Звужуючи задачу до використання жестів в ОЗ, для їх широкого застосування в системі безконтактної ЛМВ можна виділити наступні обмеження, які необхідно подолати:

1) Положення рук хірурга обмежене віртуальним вікном стерильності, що складається з рухів руками не нижче стегон і не вище рівня плечей.

2) Технологічні складнощі розпізнавання жестів, що виникають через оклюзії, неоднорідність сцени, строкатий фон і швидкі рухи, що викликають розмитість жесту на відео.

3) Відео жестів руки містить різні її форми з-за зміни положення руки під різними кутами, різні кольори хірургічних рукавичок та розмір рук хірургів.

В умовах ОЗ дуже важлива швидкість реагування системи безконтактної людино-машинної взаємодії, особливо в умовах неочікуваного та непередбачуваного розвитку подій під час проведення операції.

Також, однією з вимог до систем безконтактної ЛМВ для управління в ОЗ з використанням жестів рук є точність, яка передбачає точність розпізнавання та точність прогнозування. Критерії, що впливають на продуктивність ЛМВ, засновані на тому, наскільки близькі траєкторії та маска жесту руки до вивчених шаблонів, на основі метрик відстані, і вказують на рівень схожості або відмінності жесту від інших жестів, що використані для навчання моделей.

Головною вимогою візуального розпізнавання жестів до їх використання в інтерфейсах ЛМВ є безперервне розпізнавання в режимі реального часу з мінімальною затримкою на виході розпізнавання, але досягти такої продуктивності системи надзвичайно складно з наступних причин: (1) початкова

і кінцева точка жесту невідома; (2) захоплений кадр має негайно пройти попередню обробку, після чого всі елементи повинні бути витягнуті та використані для класифікації, перш ніж буде захоплено наступний кадр; (3) необхідно встановити спеціальні умови щодо освітлення, фону, розташування і людини і камери.

1.6 Постановка наукового завдання та обґрунтування методики досліджень

Вхідними даними для системи ЛМВ для безконтактного управління переглядом медичних зображень є відеопослідовності кадрів, що містять просторово-часові ознаки. В результаті аналізу визначено наступні обмеження, що впливають на якість розпізнавання та прогнозування жестів в системах ЛМВ:

- жести різняться за спрямованістю та формою від людини до людини, отже, існує певна нелінійність;
- дані, взяті з траєкторії будь-якого жесту, характеризуються значною варіативністю, обумовлену відмінностями у частоті дискретизації, помилками розпізнавання та наявністю шумів зображення;
- сегментація жестів рук і вилучення ознак з кадрів відеопослідовності ускладнена через просторово-часові варіації та коартикуляцію послідовних жестів;
- моделі розпізнавання та прогнозування жестів мають високу помилку узагальнення на нових даних, яка збільшується при перенавчанні моделі.
- людино-машинна взаємодія в ОЗ з використанням кінцевих автоматів потребує високої швидкості реагування та попередження помилкових спрацьовувань скінченного автомату.

Разом з тим, аналіз показав, що розглянуті архітектури здатні враховувати просторово-часові дані і дозволяють виконувати вилучення ознак, розпізнавання та прогнозування жестів для реалізації безконтактної ЛМВ і маніпулювання медичними зображеннями.

За результатами аналізу встановлено, що перспективним напрямком є розробка елементів інформаційної технології ЛМВ на основі комп'ютерного зору, таких як моделі розпізнавання та метод прогнозування для безконтактного управління переглядом медичних зображень в операційній залі з використанням жестів рук.

1.6.1 Загальне наукове завдання

Метою дисертаційної роботи є розробка нової інформаційної технології людино-машинної взаємодії для безконтактного маніпулювання медичними зображеннями в операційній кімнаті, що дозволить підвищити точність та швидкість розпізнавання та прогнозування жестів хірурга.

Результат рішення наукового завдання дозволить:

1) Підвищити ефективність систем розпізнавання статичних і динамічних жестів рук, що розбудовуються на основі комп'ютерного зору.

2) Зменшити час затримки на розпізнавання жестів і, тим самим покращити якість ЛМВ.

1.6.2 Часткові завдання досліджень

Для досягнення поставленої мети в роботі мають бути вирішені наступні завдання:

1) провести аналіз предметної області, вимог і передумов до розробки і використання систем ЛМВ за допомогою жестів;

2) розробити методологію проектування і спільного використання візуальних методів розпізнавання жестів рук;

3) розробити моделі комп'ютерного зору на основі глибоких нейронних мереж, що забезпечують відстеження та розпізнавання заданих жестів рук, здатних функціонувати в режимі реального часу, інваріантних до афінних

перетворень і зміни освітлення;

4) удосконалити модель статичного розпізнавання жестів;

5) удосконалити технологію розпізнавання та прогнозування динамічних жестів на основі моделі генерації послідовностей з використанням ConvLSTM;

6) створити новий датасет і виконати концептуальну розробку інтуїтивного словникового запасу динамічних жестів, адаптованих до особливостей хірургічного контексту;

7) удосконалити технологію взаємодії між хірургом та системою перегляду зображень в операційній залі яка використовує функції просторового домену для відстеження потенційних напрямків руху рук за допомогою скінченного автомату;

8) розробити елементи інформаційної технології для проектування системи ЛМВ з використанням жестів;

9) виконати тестування та практичне впровадження розроблених моделей, засобів і технологій.

1.6.3 Обґрунтування методики досліджень

Дисертаційне дослідження є прикладним, описовим і експериментальним.

Наукова проблема – недосконаліть методів розпізнавання жестів рук, що використовуються в системах людино-машинної взаємодії на основі комп’ютерного зору.

Область досліджень – методи машинного навчання, зокрема глибоке навчання, цифрова обробка зображень, розпізнавання образів та методи штучного інтелекту.

Методика досліджень ґрунтується на застосуванні системного підходу до вирішення загальної і часткових завдань дисертаційного дослідження і складається з детального аналізу та вибору математичного апарату, тестування базових моделей, створення та тестування власних моделей та їх порівняння з базовими, створення програмних засобів для реалізації інформаційної технології.

Методика досліджень розробляється виходячи з вимог до логіки та послідовності вирішення поставлених в роботі завдань та поділена на чотири основних етапи. На рис. 1.14 наведена запропонована в дисертаційній роботі методика досліджень, означені основні етапи, показані взаємозв'язки з отриманими результатами.

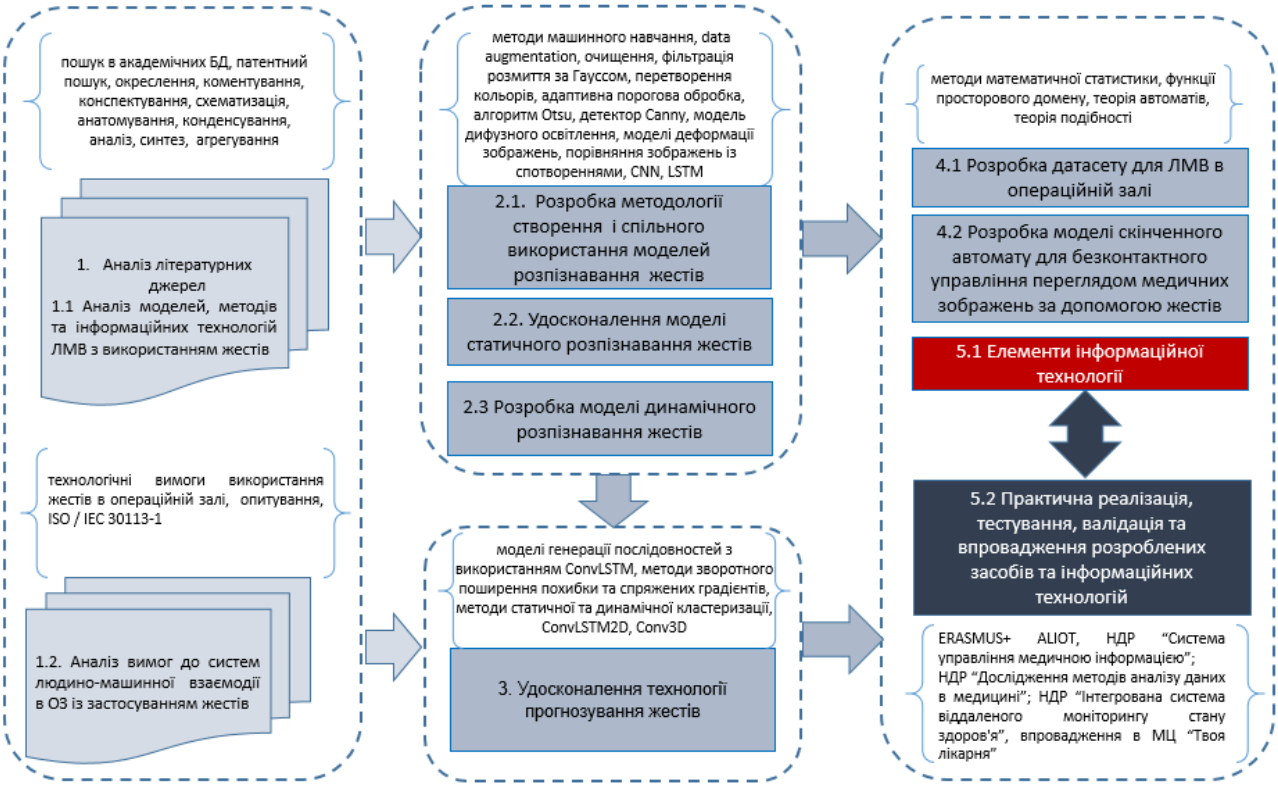


Рисунок 1.14 – Загальна схема методики дослідження

На першому етапі необхідно провести аналіз літературних джерел, наявних методів і технологій, що використовуються для розпізнавання та прогнозування жестів на підставі відео, визначити особливості побудови і функціонування мережевих архітектур, що відповідають вимогам застосування таких систем. Особлива увага приділяється аналізу методів, технологій, архітектур, що засновані на аналізі просторово-часових даних в режимі реального часу.

На другому етапі виконується розробка методології створення і спільного використання моделей розпізнавання жестів і, на її основі, – комплексу моделей для статичного та динамічного розпізнавання жестів.

На третьому етапі комплекс моделей використовується для вдосконалення технології розпізнавання та прогнозування жестів, що на виході дає послідовність прогнозованих кадрів жестів, які в свою чергу, є вхідними даними для моделі скінченного автомату для швидкої та більш точної людино-машинної взаємодії.

На останньому етапі створюється власний датасет і проводиться оцінювання розроблених та удосконалених елементів інформаційної технології та імплементації прототипу для безконтактного маніпулюваннями переглядом зображень в умовах операційної зали.

Висновки до першого розділу

1. У розділі проведено порівняльний аналіз сучасних підходів к вирішенню задачі розпізнавання статичних и динамічних жестів на основі комп'ютерного зору для реалізації людино-машинної взаємодії, зокрема розглянуто моделі, засновані на застосуванні архітектур нейронних мереж CNN і LSTM, що складають основу моделей розпізнавання статичних і динамічних жестів.

2. Виконано аналіз сучасного стану інформаційних технологій у сфері розпізнавання жестів, виявлено їх переваги та недоліки, складено таблиці з результатами їх роботи.

3. За результатами проведеного аналізу і порівняння існуючих рішень зроблено висновок про актуальність дисертаційної роботи. Сформульоване загальне науково-прикладне завдання роботи, яке полягає в розробці засобів розпізнавання та прогнозування жестів рук для реалізації людино-машинної взаємодії в умовах операційної зали. Визначені задачі досліджень та виявлені основні вимоги до вирішення проблеми розпізнавання жестів на відеопослідовністях у режимі реального часу.

4. Визначено, що існує необхідність розвитку технологій розпізнавання жестів, які забезпечать основу функціонування стерильної, інтуїтивно зрозумілої

інтерактивної системи навігації зображеннями за допомогою жестів рук в операційній залі.

5. Основна мета дослідження полягає покращенні характеристик автоматичного розпізнавання статичних та динамічних жестів рук на зображеннях, у відеопослідовностях та в режимі реального часу за рахунок розробки та практичного використання моделей і методу інформаційної технології людино-машинної взаємодії з використанням жестів

Результати розділу опубліковано в роботах автора [2, 3, 7, 8, 9, 14] (Додаток А)

Література до першого розділу

1. Багрій Р.О. Інформаційна технологія альтернативної комунікації для людей з обмеженими можливостями спілкування: автореф. дис. ... канд. техн. наук : 05.13.06. Тернопіл. нац. екон. ун-т. Тернопіль, 2018. 24 с.

2. Бармак О.В. Моделювання та аналіз невербальних каналів комунікації для створення нових інформаційних технологій.- Дисертація д-ра техн. наук: 05.13.06, Нац. акад. наук України, Ін-т кібернетики ім. В. М. Глушкова. К, 2013. 203 с.

3. Бармак О.В., Багрій Р.О., Крак Ю.В., Касьянюк В.С. Інформаційна технологія альтернативної комунікації для людей з обмеженими можливостями спілкування. *Штучний інтелект,* 2018. № 2. С. 16-24. Режим доступу: http://nbuv.gov.ua/UJRN/II_2018_2_4.

4. Давидов М.В. Математичне та програмне забезпечення комп'ютерної системи ідентифікації елементів української жестової мови : дис. ... канд. техн. наук : 01.05.03; Нац. ун-т «Львів. політехніка». Л., 2009. 162 с.

5. Кондратенко Ю.П., Доценко А.С. (2011) Нечітка логіка в задачах розпізнавання жестів. *Наукові праці Чорноморського державного університету імені Петра Могили. Сер.: Комп'ютерні технології*, № 160 Вип. 148, С. 74-79.

6. Сергієнко І.В., Крак Ю.В., Бармак О.В, Куляс А.І. Системи жестової комунікації: моделювання та розпізнавання дактильної жестової мови. Монографія. К: Наукова думка, 2019. 284 с.

7. Alashhab S., Gallego A.-J., Lozano M.A. (2018) Hand Gesture Detection with Convolutional Neural Networks, *Advances in Intelligent Systems and Computing*, pp. 45-52.

8. Balbin J.R., Padilla D.A., Caluyo F.S., Fausto J.C., Hortinela C.C., Manlises C.O., Bernardino C.K.S., Fiñones E.G., Ventura L.T. (2016) Sign language word translator using Neural Networks for the Aurally Impaired as a tool for communication. In *2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE).* pp. 425-429.

9. Barros P., Maciel-Junior N.T., Fernandes B.J., Bezerra B.L., Fernandes S.M. (2017) A dynamic gesture recognition and prediction system using the convexity approach. *Computer Vision and Image Understanding*, vol. 155, pp. 139-149.

10. Bheda V., Radpour D. (2017) Using deep convolutional networks for gesture recognition in American sign language. *arXiv preprint arXiv:1710.06836*.

11. Bhuiyan M., Picking R. (2011) A Gesture Controlled User Interface for Inclusive Design and Evaluative Study of Its Usability, *Journal of Software Engineering and Applications*, Vol. 4 No. 9, 2011, pp. 513-521. doi: 10.4236/jsea.2011.49059.

12. Bourke A.K., O'Brien J.V., Lyons G. M. (2007) Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm. *Gait & posture* 26.2, pp. 194-199.

13. Buxton H. (2003) Learning and understanding dynamic scene activity: a review. *Image and Vision Computing*, vol. 21(1), pp. 125-136.

14. Camastra F., Vinciarelli A. Machine Learning for Audio, Image and Video Analysis: Theory and Applications. *Advanced Information and Knowledge Processing*. 2nd ed. 2015, 577 p.

15. Castrejon L., Ballas N., Courville A. (2019) Improved Conditional VRNNs for Video Prediction," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, pp. 7607-7616, doi: 10.1109/ICCV.2019.00770.

16. Chang C.C., Chen J.J., Tai W., Han C.C. (2006) New approach for static gesture recognition. J Inform Sci Eng., vol. 22, pp. 1047-1057.

17. Chen Q., Xie Q., Yuan, Huang H., Li Y. (2019). Research on a Real-Time Monitoring Method for the Wear State of a Tool Based on a Convolutional Bidirectional LSTM Model. *Symmetry*, 11(10), 1233. doi:10.3390/sym11101233

18. Cheok M.J., Omar Z., Jaward M.H. (2019) A review of hand gesture and sign language recognition techniques. *Int. J. Mach. Learn. & Cyber*. 10, pp. 131-153. https://doi.org/10.1007/s13042-017-0705-5

19. Cui R., Cao Z., Pan W., Zhang C., Wang, J. (2020). Deep Gesture Video Generation with Learning on Regions of Interest. In *IEEE Transactions on Multimedia*, 22, pp. 2551-2563. doi:10.1109/tmm.2019.2960700

20. Dadashzadeh A., Alireza T.T., Tahmasbi M. (2018) HGR-Net: A Two-stage Convolutional Neural Network for Hand Gesture Segmentation and Recognition. *ArXiv abs/1806.05653* (2018).

21. Davis J., Shah M. (1994) Visual gesture recognition. *Vision, Image and Signal Processing*, vol. 141(2), pp. 101-106.

22. Devineau G., Moutarde F., Xi W., Yang J. (2018) Deep Learning for Hand Gesture Recognition on Skeletal Data. *13th IEEE Conference on Automatic Face and Gesture Recognition (FG'2018)*, pp. 106-113.

23. Donahue J., Hendricks L. A., Guadarrama S., Rohrbach M., Venugopalan S., Saenko K., Darrell T. (2015) Long-term recurrent convolutional networks for visual recognition and description, in *CVPR*, 2015.

24. Forbes (2019) Open Innovation in Japan Breaks New Ground In The Operating Room. https://www.forbes.com/sites/japan/2019/03/08/open-innovation-in-japan-breaks-new-ground-in-the-operating-room/?sh=589f058714a8 (08.11.2020).

25. Freeman W., Tanaka K., Ohta J., Kyuma K. (1996) Computer Vision for Computer Games. *Tech. Rep. and International Conference on Automatic Face and Gesture Recognition*,

26. Fukushima T., Miyazaki F., Nishikawa A. (2012) The Development of a Noncontact Letter Input Interface “Fingual” Using Magnetic Dataset. *Transactions of the Society of Instrument and Control Engineers* 48.3, pp. 159-166.

27. Gallo L., Placitelli A.P., Ciampi M. (2011) Controller-free exploration of medical image data: Experiencing the Kinect. In *Proceedings of the 2011 24th international symposium on computer-based medical systems (CBMS),* Bristol, UK, 27–30 June 2011; pp. 1-6.

28. Gandy M., Starner T., Auxier J., Ashbrook D. (2000) The Gesture Pendant: A Self Illuminating, Wearable, Infrared Computer Vision System for Home Automation Control and Medical Monitoring. *Proc. of IEEE Int. Symposium on Wearable Computers*, pp. 87-94.

29. Gope D.C. (2011) Hand Gesture Interaction with Human-Computer. *Global Journal of Computer Science and Technology*, vol. 11(23).

30. Goza S. M., Ambrose R. O., Diftler M.A., Spain I.M. (2004) Telepresence Control of the NASA/DARPA Robonaut on a Mobility Platform. In: *Proc of the 2004 Conference on Human Factors in Computing Systems. ACM Press*, pp. 623-629.

31. Graetzel C., Fong T.W., Grange C., Baur C. (2004) A non-contact mouse for surgeon-computer interaction. *Technology and Health Care* 12, 3 (Aug. 24, 2004), pp. 245–257.

32. Gupta N., Mittal P., Dutta Roy S., Chaudhury S., Banerjee S. (2002) Developing a Gesture-Based Interface. *IETE Journal of Research*, vol. 48(3), pp. 237-244.

33. Hochreiter S., J. Schmidhuber (1997) Long short-term memory. *Neural computation* 9.8, pp. 1735-1780.

34. Hong P., Turk M., Huang T.S. (2000) Gesture modeling and recognition using finite state machines. In *Proceedings of the Fourth IEEE International*

Conference on Automatic Face and Gesture Recognition, Grenoble, France, 28–30 March 2000; pp. 410-415.

35. Huang Z., Leng J. (2010). Analysis of Hu's moment invariants on image scaling and rotation. *2010 2nd International Conference on Computer Engineering and Technology*. doi:10.1109/iccet.2010.5485542.

36. Hussain S., Saxena R., Han X., Khan J. A., and Shin H. (2017) Hand gesture recognition using deep learning." *2017 International SoC Design Conference (ISOCC)*, pp. 48-49.

37. ISO/IEC 30113-11:2017(en) Information technology – Gesture-based interfaces across devices and methods – Part 11: Single-point gestures for common system actions. Режим доступу: https://www.iso.org/obp/ui/#iso:std:iso-iec:30113:-11:ed-1:v1:en (21.11.2020).

38. ISO/IEC 30113-12:2019(en) Information technology – User interfaces – Gesture-based interfaces across devices and methods — Part 12: Multi-point gestures for common system actions. Режим доступу: https://www.iso.org/obp/ui/#iso:std:iso-iec:30113:-12:ed-1:v1:en (21.11.2020).

39. ISO/IEC 30113-60:2020(en) Information technology – Gesture-based interfaces across devices and methods – Part 60: General guidance on gestures for screen readers. Режим доступу: https://www.iso.org/obp/ui/#iso:std:iso-iec:30113:-60:ed-1:v1:en (21.11.2020).

40. Jing L., Vahdani E., Huenerfauth M., Tian Y. (2019) Recognizing American Sign Language Manual Signs from RGB-D Videos. *arXiv preprint* arXiv:1906.02851.

41. John V., Umetsu M., Boyali A., Mita S., Imanishi M., Sanma N., Shibata S. (2017) Real-Time Hand Posture and Gesture-Based Touchless Automotive User Interface Using Deep Learning, *IEEE Intelligent Vehicles Symposium (IV)*, pp. 869-874.

42. Juang C.-F., Ku K.-C. (2005) A Recurrent Fuzzy Network for Fuzzy Temporal Sequence Processing and Gesture Recognition, *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 35, no. 4, pp. 646-658.

43. Kanokoda T., Kushitani Y., Shimada M., Shirakashi J. I. (2019). Gesture Prediction Using Wearable Sensing Systems with Neural Networks for Temporal Data Analysis. *Sensors (Basel, Switzerland)*, 19(3), 710. https://doi.org/10.3390/s19030710

44. Khan R.Z., Ibraheem, N.A. (2012). Hand gesture recognition: a literature review. International journal of artificial Intelligence & Applications, 3(4), p.161.

45. Konrad T., Demirdjian D., Darrell T. (2003) Gesture + Play: Full-Body Interaction for Virtual Environments. In: *CHI '03 Extended Abstracts on Human Factors in Computing Systems*. ACM Press, 620–621.

46. Köpüklü O., Gunduz A., Kose N., Rigoll G. (2019) Real-time Hand Gesture Detection and Classification Using Convolutional Neural Networks, *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, Lille, France, 2019, pp. 1-8, doi: 10.1109/FG.2019.8756576.

47. Larsen A.B.L., Sønderby S.K., Larochelle H., Winther O. (2015) Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint* arXiv:1512.09300, 2015.

48. Lee A.X., Zhang R., Ebert F., Abbeel P., Finn C., Levine S. (2018) Stochastic adversarial video prediction. arXiv preprint arXiv:1804.01523, 2018.

49. Lee A., Cho Y., Jin S., Kim N. (2020). Enhancement of surgical hand gesture recognition using a capsule network for a contactless interface in the operating room. *Computer Methods and Programs in Biomedicine*, 190, 105385. doi:10.1016/j.cmpb.2020.105385.

50. Lei Q., Zhang H., Xia Z., Yang Y., He Y., Liu S. (2018) Applications of hand gestures recognition in industrial robots: a review , *Proc. SPIE 11041, Eleventh International Conference on Machine Vision (ICMV 2018), 110411Q* (15 March 2019); https://doi.org/10.1117/12.2522962

51. Li Y. et al. (2018) Deep attention network for joint hand gesture localization and recognition using static RGB-D images. *Information Sciences 441* pp. 66-78.

52. Li G. et al. (2019) Hand gesture recognition based on convolution neural network. *Cluster Computing* 22.2: pp. 2719-2729.

53. Li D., Yu X., Xu C., Petersson L., Li H. (2020) Transferring cross-domain knowledge for video sign language recognition. *arXiv preprint* arXiv:2003.03703.

54. Liao Y., Xiong P., Min W., Min W. and Lu J. (2019) Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks. *IEEE Access*, 7, pp.38044-38054.

55. Lin H.-I., Hsu M.-H., Chen W.-K. (2014) Human hand gesture recognition using a convolution neural network. *Automation Science and Engineering (CASE), 2014 IEEE International Conference on*. IEEE, 2014.

56. Lotter W., Kreiman G., Cox D. (2016) Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint* arXiv:1605.08104

57. Luo Q. et al. (2010) Human action detection via boosted local motion histograms. *Machine Vision and Applications* 21.3 (2010): 377-389.

58. Molchanov P., Gupta S., Kim K., 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.

59. Molchanov P., Gupta S., Kim K., Pulli K. (2015) Multi-sensor system for driver's hand-gesture recognition. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference*, vol. 1, pp. 1-8.

60. Molchanov P., Yang X., Gupta S., Kim K., Tyree S., Kautz J. (2016) Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4207-4215.

61. Moni M.A., Ali A.B.M.S. (2009) HMM based hand gesture recognition: A review on techniques and approaches. *IEEE International Conference on Computer Science and Information Technology*, Beijing, China, 2009, pp. 433-437, doi: 10.1109/ICCSIT.2009.5234536.

62. Nair V., Clark J.J. (2002) Automated visual surveillance using Hidden Markov models, In *International Conference on Vision Interface*, pp. 88-93.

63. Nasri N., Orts-Escolano S., Gomez-Donoso F., Cazorla M. (2019) Inferring Static Hand Poses from a Low-Cost Non-Intrusive sEMG Sensor. *Sensors 2019*, vol. 19, p. 371.

64. New J.R., Hasanbelliu E., Aguilar M. (2003) Facilitating User Interaction with Complex Systems via Hand Gesture Recognition. In *Proc. of Southeastern ACM Conf., Savannah*, (2003).

65. Nielsen M., Störring M., Moeslund T.B., Granum E. (2004) A Procedure for Developing Intuitive and Ergonomic Gesture Interfaces for HCI. In: *Camurri A., Volpe G. (eds) Gesture-Based Communication in Human-Computer Interaction. GW 2003. Lecture Notes in Computer Science*, vol 2915. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24598-8_38.

66. Nishida N., Nakayama H. (2016) Multimodal Gesture Recognition Using Multi-stream Recurrent Neural Network, *Image and Video Technology Lecture Notes in Computer Science*, pp. 682-694.

67. Nishikawa A. et al. (2003) FAce MOUSe: A novel human-machine interface for controlling the position of a laparoscope, in *IEEE Transactions on Robotics and Automation*, vol. 19, no. 5, pp. 825-841, Oct. 2003, doi: 10.1109/TRA.2003.817093.

68. Nogales R., Benalcázar M.E. (2020) A Survey on Hand Gesture Recognition Using Machine Learning and Infrared Information. In: *Botto-Tobar M., Zambrano Vizuete M., Torres-Carrión P., Montes León S., Pizarro Vásquez G., Durakovic B. (eds) Applied Technologies. ICAT 2019. Communications in Computer and Information Science*, vol. 1194. Springer, Cham. https://doi.org/10.1007/978-3-030-42520-3_24

69. O'Hara K., Dastur N., Carrell T., Gonzalez G., Sellen A., Penney G., … Rouncefield M. (2014). Touchless interaction in surgery. *Communications of the ACM*, vol. 57(1), pp. 70-77. doi:10.1145/2541883.2541899

70. Ohn-Bar E., Trivedi M. M. (2014) Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE transactions on intelligent transportation systems*, vol. 15(6), pp. 2368-2377.

71. OR2020: Operating room of the future Workshop, 2004 https://apps.dtic.mil/sti/pdfs/ADA430482.pdf (23.10.2020)

72. Oudah M., Al-Naji A., Chahl J. (2020) Hand Gesture Recognition Based on Computer Vision: A Review of Techniques. *Journal of Imaging, 6(8), 73*. doi:10.3390/jimaging6080073

73. Parvathy P., Subramaniam K., Prasanna V.G.K.D. et al. (2020) Development of hand gesture recognition system using machine learning. *J Ambient Intell Human Comput* . https://doi.org/10.1007/s12652-020-02314-2.

74. Parvini F, Shahabi C., 2007. An algorithmic approach for static and dynamic gesture recognition utilizing mechanical and biocharacteristics. Int J Bioinform Res Appl, Vol. 3, Issue 1.

75. Pigou L., Dieleman S., Kindermans P.-J., Schrauwen B. (2015) Sign language recognition using convolutional neural networks. *Computer Vision - ECCV 2014 Workshops. Lecture Notes in Computer Science*. Springer, pp. 572-578.

76. Rauschert I., Agrawal P., Sharma R., Fuhrmann S., Brewer I., MacEachren A. (2002) Designing a human-centered, multimodal GIS interface to support emergency management. In *Proceedings of the 10th ACM international symposium on Advances in geographic information systems.* pp. 119-124.

77. Rautaray S.S., Agrawal A. (2012) Vision Based Hand Gesture Recognition for Human Computer Interaction: A survey, *Springer Transaction on Artificial Intelligence Review*, pp. 1-54.

78. Sánchez-Margallo F.M., Sánchez-Margallo J.A., Pagador J.B., Moyano J.L., Moreno J., Usón J. (2010) Ergonomic Assessment of Hand Movements in Laparoscopic Surgery Using the CyberGlove®. In: Miller K., Nielsen P. (eds) Computational Biomechanics for Medicine. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-5874-7_13

79. Sarkar A.R, Sanyal G., Majumder S. (2013) Hand gesture recognition systems: a survey. *International Journal of Computer Applications* 71.15.

80. Sharma R., Huang T. S., Pavovic V. I., Zhao Y., Lo Z., Chu S., Schulten K., Dalke A., Phillips J., Zeller M., Humphrey W. (1996) Speech/Gesture Interface to

a Visual Computing Environment for Molecular Biologists". In: Proc. of ICPR'96 II (1996), 964-968.

81. Shi X., Chen Z., Wang H., Yeung D.-Y., Wong W.-k., Woo W.-c. (2015) Convolutional LSTM Network: A machine learning approach for precipitation nowcasting. In *Proceedings of the 28th International Conference on Neural Information Processing Systems* – Vol. 1 (NIPS'15). MIT Press, Cambridge, MA, USA, pp. 802–810.

82. Sihombing P., Muhammad R. B., Herriyance H. and Elviwani E. (2020) Robotic Arm Controlling Based on Fingers and Hand Gesture, *2020 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT)*, Medan, Indonesia, 2020, pp. 40-45, doi: 10.1109/MECnIT48290.2020.9166592.

83. Simonyan K., Zisserman A. (2014) Two-Stream Convolutional Networks for Action Recognition in Videos, *arXiv preprint arXiv:1406.2199*.

84. Smith G. M. & Schraefel. M. C. (2004) The Radial Scroll Tool: Scrolling Support for Stylus-or Touch-Based Document Navigation". In *Proc. 17th ACM Symposium on User Interface Software and Technology*. ACM Press, (2004) 53–56.

85. Starner T., Weaver J., Pentland A. (1998) Real-Time American Sign Language Recognition using Desk and Wearable Computer Based Video. *PAMI*, vol. 20(12), pp. 1371-1375.

86. Stergiopoulou E., Papamarkos N. (2009) Hand gesture recognition using a neural network shape fitting technique. *Engineering Applications of Artificial Intelligence* 22.8 (2009): 1141-1158.

87. Stotts D., Smith J.M., Gyllstrom K. (2004) Facespace: Endo- and Exo-Spatial Hypermedia in the Transparent Video Facetop". In: *Proc. of the Fifteenth ACM Conf. on Hypertext & Hypermedia. ACM Press*, (2004) 48–57.

88. Tang G. (2016) The Development of a Human-Robot Interface for Industrial Collaborative System. PhD thesis. – Cranfield University. School of Engineering Aero-structure Assembly and Systems Installation Group http://dspace.lib.cranfield.ac.uk/handle/1826/10213.

89. Tang Y., Zou W., Jin Z., Li X. (2018) Multi-Scale Spatiotemporal Conv-LSTM Network for Video Saliency Detection. *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval - ICMR ’18,* pp. 362-369. https://doi.org/10.1145/3206025.3206052

90. Tomasi C., Petrov S., Sastry A. (2003) 3d tracking = classification+ interpolation. *null*. IEEE, 2003.

91. Trigueiros P., Ribeiro F., Reis L.P. (2012) A comparison of machine learning algorithms applied to hand gesture recognition," *7th Iberian Conference on Information Systems and Technologies (CISTI 2012)*, Madrid, Spain, 2012, pp. 1-6.

92. Tyagi A., Bansal S. (2021) Feature Extraction Technique for Vision-Based Indian Sign Language Recognition System: A Review. In: *Singh V., Asari V., Kumar S., Patel R. (eds) Computational Methods and Data Engineering. Advances in Intelligent Systems and Computing*, vol 1227. Springer, Singapore. https://doi.org/10.1007/978-981-15-6876-3_4

93. Wachs J.P., Kölsch M., Stern H., Edan Y. (2011) Vision-based hand-gesture applications, *Communications of the ACM*, v.54, no.2, pp. 60-71.

94. Wang S., et al. (2016) Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 851-860.

95. Wang X., Xia M., Cai H., Gao Y., Cattani C. (2012) Hidden-Markov-Models-Based Dynamic Hand Gesture Recognition, *Mathematical Problems in Engineering*, vol. 2012, Article ID 986134, 11 p.

96. Wei Y., Lu Y., Easterbrook S., Fidler S. (2019) Efficient and information-preserving future frame prediction and beyond. *International Conference on Learning Representations*. Режим доступу: https://openreview.net/pdf?id=B1eY_pVYvB (16.08.2020).

97. Wu X., Wu Z., Zhang J., Ju L.,Wang S. (2020) SalSAC: A Video Saliency Prediction Model with Shuffled Attentions and Correlation-based ConvLSTM. In *AAAI*, pp. 12410-12417.

98. Xiong F., Shi X., Yeung D.Y. (2017) Spatiotemporal modeling for crowd counting in videos. In *Proceedings of the IEEE International Conference on Computer Vision,* pp. 5151-5159

99. Yamato J., Ohya J., Ishii K. (1992) Recognizing human action in time-sequential images using hidden Markov model. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Proceedings CVPR'92.

100. Yang M.H., Ahuja N., Tabb M. (2002) Extraction of 2-D Motion Trajectories and its Application to Hand Gesture Recognition," in *PAMI,* vol. 29(8), pp. 1062-1074.

101. Yasen M, Jusoh S. (2019). A systematic review on hand gesture recognition techniques, challenges and applications. *Peer J. Computer Science* 5:e218 https://doi.org/10.7717/peerj-cs.218

102. Yeasin M., Chaudhuri S. (2000) Visual understanding of dynamic hand gestures. *Pattern Recognit*, 33, pp. 1805-1817.

103. Young-Guk K., Quan Le M. (2017) DeepForest: 3D Hand Pose Estimation Using Deep Network and Random Forest Regression. In: *The AAAI-17 Workshop on Crowdsourcing, Deep Learning, and Artificial Intelligence Agents WS-17-07*.

104. Zeng J., Sun Y., Wang F. (2012) A natural hand gesture system for intelligent human-computer interaction and medical assistance. In *Proceedings of the 2012 Third Global Congress on Intelligent Systems*, Wuhan, China, 6–8 November 2012, pp. 382-385.

105. Zhang Z., He C., Yang K. (2020) A Novel Surface Electromyographic Signal-Based Hand Gesture Prediction Using a Recurrent Neural Network. *Sensors* 2020, *20*, 3994. https://doi.org/10.3390/s20143994

РОЗДІЛ 2
РОЗРОБЛЕННЯ МЕТОДОЛОГІЇ РОЗРОБКИ МОДЕЛЕЙ РОЗПІЗНАВАННЯ ЖЕСТІВ

У розділі пропонується нова методологія розробки моделей розпізнавання статичних та динамічних жестів рук за допомогою CNN та CNN+LSTM. Розкрито етапи розробки моделей, процедури збільшення набору даних, очищення даних, зміну колірної моделі, сегментацію та виділення контуру під час попередньої обробки, що сприяє кращому вилученню ознак, застосування операції згортки квадратною матрицею з непарною розмірністю. Розкрито математичний апарат формування ознак та класифікації у згортковій нейронній мережі.

2.1 Постановка задачі розпізнавання жестів

Процес розпізнавання будь-якого жесту, статичного або динамічного, передбачає наявність в системі ЛМВ даних про *повністю виконаний* жест.

У загальному виді, завдання розпізнавання жестів рук полягає в розробці класифікатора для розпізнавання різних жестів рук, де кожен жест має значення одного простого або складеного слова. У контексті реалізації ЛМВ, метою є створення простої, дешевої і ефективної по ресурсоємності моделі, здатної ефективно використовувати вбудовані в неї класифікатори і оперувати в різних середовищах відносно освітлення та кутів обертання рук з мінімальною затримкою.

Оскільки основою системи ЛМВ є класифікатори жестів, далі, задача розпізнавання формулюється в контексті розпізнавання статичних та динамічних жестів.

2.1.1 Постановка задачі розпізнавання статичних жестів

Нехай g_i- статичний жест, а $\mathbb{S} = \{S_i\}_{i=1}^{n}$ набір класів статичних жестів, де n визначає кількість класифікованих жестів. Нехай також дана послідовність

вхідних зображень $X = x_1, \ldots, x_t, \ldots, x_T$, де кожне зображення містить жест g_i класу i, тобто для кожного зображення відомий клас жесту, тоді завдання розпізнавання статичних жестів полягає у присвоєнні відповідного позначення кадру за модальністю кожному x_t. Тобто завдання класифікації жесту може бути сформульовано як задача пошуку приналежності зображення X_t до певного класу жестів S_i, тобто знайти пару (X_t, S_i), розподіл ймовірності $P(X_t, S_i)$ має максимальне значення $\forall i$:

$$P(X_i, S_i) \rightarrow \max. \tag{2.1}$$

Припустимо, що система ЛМВ отримує серію розпізнаних жестів $\boldsymbol{\mathcal{G}}$ від класифікатора, протягом періоду відповіді, тобто

$$\mathcal{G} = \{g_1, g_2, \ldots, g_n\}, \tag{2.2}$$

де g_i - це i-й жест, розпізнаний системою.

Тоді, дискримінантна функція для виконання команди k_i, яка відповідає жесту g_i, може бути представлена у вигляді ймовірнісної моделі (2.3):

$$f_i(\mathcal{G}) = Pr(k_i|\mathcal{G}). \tag{2.3}$$

Для знаходження f_i можна використовувати різні підходи - моделі Маркова, оцінку ризику Байєса та ін. Однак, у реальному середовищі, статистичні дані про ризик і частоту появи кожного жесту та команди можуть бути не доступні. У цьому випадку можна представити дискримінантну функцію (2.3) у спрощеному вигляді [24]:

$$f_i(\mathcal{G}) = \frac{\sum_j w_{i,j}}{|\mathcal{G}|}, \tag{2.4}$$

де

$$w_{i,j} = \begin{cases} 1, & g_j = k_i \\ 0, & g_j \neq k_i \end{cases}, \tag{2.5}$$

де $j = 1,\ldots, n$ та $|\cdot|$ є операцією потужності.

Модель (2.4) використовує поняття „період відповіді”, що фактично означає, що система ЛМВ не виводить відповідь одразу після того, як отримує жест, розпізнаний класифікатором CNN, а виконує команду відповідно до дискримінантної функції, яка обчислюється за допомогою послідовних жестів, що з'являються протягом певного періоду.

Варто відзначити, що на відміну від (2.3), задача знаходження дискримінантної функції у вигляді (2.4) може ефективно покращити надійність системи ЛМВ, але також може призвести до затримки реакції системи. Засіб вирішення цієї проблеми полягає в тому, що система змушена продовжувати виконувати «затриману» дію, як тільки дія буде підтверджена, і поки нова дія не буде підтверджена схемою відповідей, обговореною вище.

2.1.2 Постановка задачі розпізнавання динамічних жестів

Нехай d_i - динамічний жест, а $\mathbb{C} = \{C_l\}_{l=1}^{N}$ набір класів динамічних жестів, де N визначає кількість класифікованих жестів. Тоді, зміни d_i в часі можна визначити як [13]:

$$D_i = \{\Im_i^t\}_{t=1}^{T_i}, \tag{2.6}$$

де $t \in [1, T_i]$ визначає певний момент у часовому вікні розміром T_i, а $\Im_i^t$ являє собою відповідний кадр d_i в момент часу t.

Враховуючи факт, що залежно від жесту або від схильності користувача, однакові жести можна виконувати з різним часом (змінним тимчасовим вікном), задача динамічної класифікації жестів рук може бути сформульована як пошук класу C_l, до якого з найбільшою ймовірністю належить жест d_i, тобто

знаходження пари (d_i, C_l), для якої розподіл ймовірності $P(d_i, C_l)$ має максимальне значення $\forall\, l$:

$$P(d_i, C_l) \rightarrow \max. \tag{2.7}$$

Нехай Φ – відображення, яке перетворює простір та часову інформацію, пов'язану з жестом d_i в єдине зображення, яке визначається як:

$$I_i = \Phi\ (D_i). \tag{2.8}$$

Припустимо, що для цього подання існує одне зображення I_i для кожного жесту d_i незалежно від розміру тимчасового вікна T_i.

Тоді завдання класифікації динамічного жесту може бути сформульовано аналогічно задачі класифікації статичного жесту, тобто як задача пошуку приналежності зображення I_i до певного класу жестів C_l, де необхідно знайти пару (I_i, C_l), розподіл ймовірності P (I_i, C_l) якої має максимальне значення $\forall\, l$:

$$P(I_i, C_l) \rightarrow \max. \tag{2.9}$$

Типи жестів визначаються на основі конкретного набору даних.

Дискримінантна функція для виконання команди k_i така сама як і у випадку роботи зі статичними жестами, тобто (2.2) або (2.3).

2.2 Методологія розробки і використання моделей розпізнавання жестів

Виходячи зі спільності постановок задач класифікації жестів, у цьому підрозділі запропоновано нову методологію розробки і спільного використання візуальних методів розпізнавання жестів рук. Пропонована методологія дозволяє створювати та досліджувати моделі для розпізнавання статичних і динамічних

жестів і складається з наступних етапів (рис. 2.1): отримання наборів даних; збільшення даних (на випадок використання власного відео та/або зображень); підготовка та попередня обробка даних; навчання класифікаторів і вилучення ознак за допомогою глибокої нейронної мережі класу CNN і, безпосередньо, розпізнавання жестів, яке може виконуватися за допомогою мереж CNN, RNN та їх комбінацій.

Етап 1. Отримання наборів вихідних даних.

Для роботи з жестами рук зазвичай можна використовувати два підходи: формування власного набору або використання відкритих наборів даних.

Етап 2. Збільшення даних.

На цьому етапі до отриманого набору даних застосовується операція збільшення даних.

Етап 3. Підготовка та попередня обробка даних.

Попередня обробка даних містить очищення даних, зміну колірної моделі, сегментацію та виявлення контуру.

Етап 4. Вилучення ознак.

Витяг ознак здійснюється шляхом операції згортки квадратною матрицею з непарною розмірністю. Матриця, яка є ядром згортки або фільтром, ковзає по зображенню, виконуючи покрокове скалярне перетворення. На виході маємо карти ознак, кількість яких залежить від числа застосовуваних фільтрів. Кожен фільтр являє собою систему поділюваних ваг і призначений для пошуку і виділення ознак за певним шаблоном.

Згортка представлена як:

$$Conv(\omega * y)_{ij} = \sigma\left(\sum_{a=0}^{m-1}\sum_{b=0}^{m-1}\omega_{ab} \times y^{l-1}_{(i+a)(j+b)}\right), \tag{2.10}$$

де ω – ядро згортки розміру m×m, y – входи з попереднього шару, σ - функція активації нейронів.

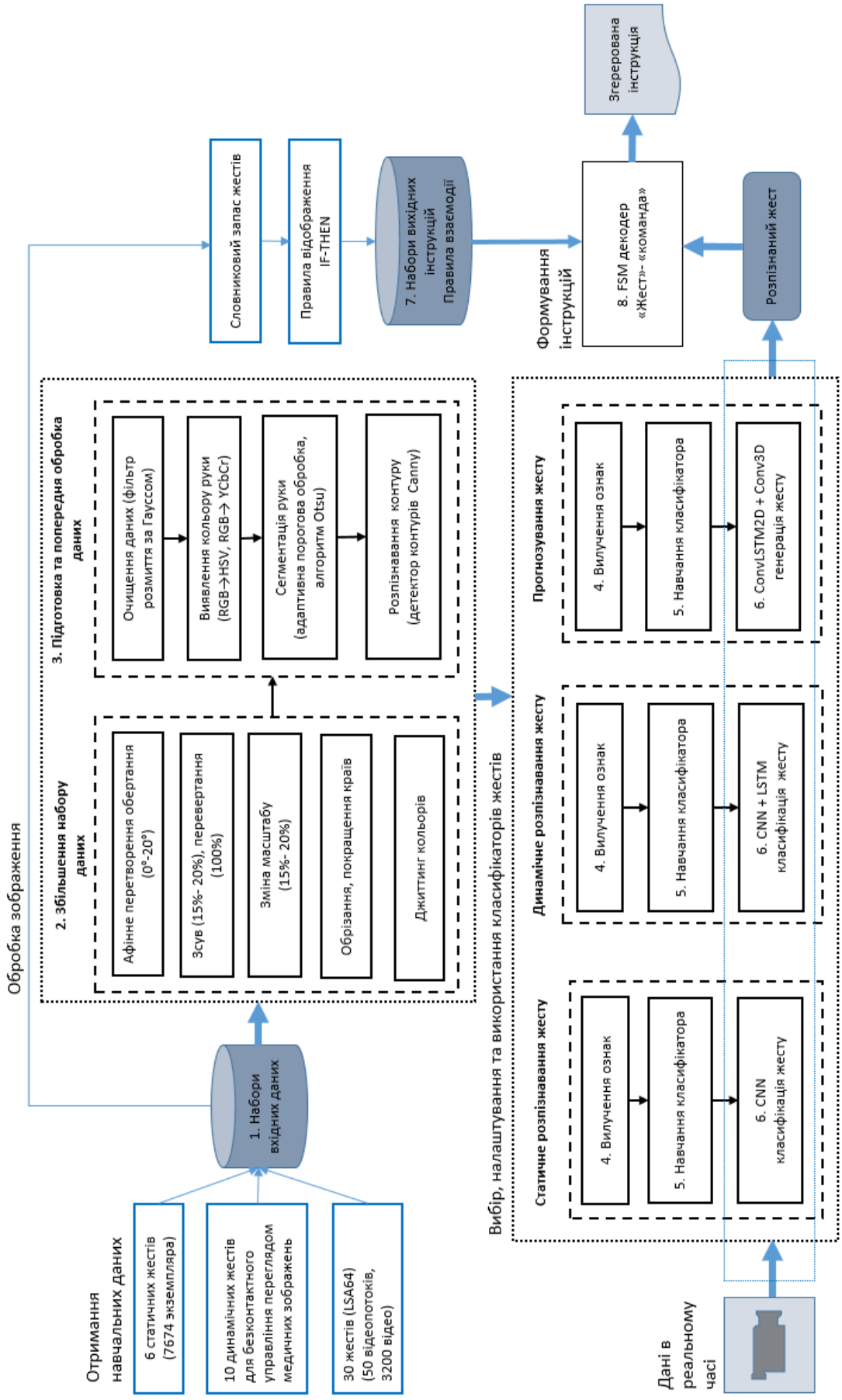


Рисунок 2.1 – Основні етапи методології розробки і спільного використання візуальних методів розпізнавання статичних і динамічних жестів рук

За кожним згортковим шаром слідує агрегувальний шар max pooling, який зменшує розмірність карти ознак, проходячи по зображенню і покроково вибираючи з рецептивного поля розміром 2×2 максимальне значення. Крім зниження розмірності в два рази і зменшення кількості параметрів в нейронної мережі, max pooling робить знайдені ознаки більш яскраво вираженими, а мережа стає більш інваріантною до місцезнаходження об'єкта на карті ознак, до зрушень і поворотів. Вихід max poolimg обчислюється за формулою (2.11).

$$y^{l+1} = \max_{0 \le i < H, 0 \le j < W} x^{l}_{i^{l+1} \times H + i, j^{l+1} \times j, d}, \tag{2.11}$$

де *H,W* – розмір вікна субдискретизації, *x* – вихідні дані.

Eman 5. Навчання класифікатора.

Частина мережі, що безпосередньо виконує класифікацію, складається зі згладжувального шару flatten, кожен вузол якого відповідає одному значенню з вектору ознак, і двох повнозв'язних шарів dense. Останній шар є вихідним і реалізує функцію softmax. Дані з останнього шару субдискретизації надходять на згладжувальний шар flatten, перетворюючись у ньому в одновимірний вектор, тобто просторові розмірності шар згортає у розмірність каналу зображення. Обчислення значень нейронів для повноз'єднаного шару відбувається за формулою (2.12).

$$x^{l}_{i} = \sum_{k=0}^{m} w^{l}_{ki} y^{l-1}_{k} + b^{l}_{i}, \tag{2.12}$$

де w^{l}_{ki} – вага від k-го нейрона шару *l*-1 до і-того нейрону поточного шару *l*, *b* – зміщення поточного шару, у – вхідні дані з попереднього шару.

Останній шар мережі з кількістю виходів що дорівнює кількості категорій що розпізнаються, реалізує нормовану експоненціальну функцію softmax. Softmax привласнює значення, представлене невід'ємним дійсним числом, кожному класу, відображаючи ймовірність приналежності. Сума всіх вихідних

сигналів дорівнює одиниці. Значення вихідного сигналу і-го нейрона відповідає ймовірності того, що правильна відповідь є і. Значення і-го виходу в softmax визначається за формулою (2.13).

$$S_i \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}, i = 1,...,n. \quad (2.13)$$

Етап 6. Розпізнавання / прогнозування жесту.

На етапі розпізнавання навчена нейронна мережа через веб-камеру в режимі реального часу розпізнає клас одержуваного на вхід жесту.

Етап 7. Створення БД вихідних інструкцій та правил взаємодії.

Етап передбачає створення словника жестів і правил відображення команд взаємодії у вигляді IF-THEN.

Етап 8. Формування інструкцій ЛМВ.

На останньому етапі для ефективного перетворення жестів в інструкції використовується детермінована модель на основі скінченного автомату (finite-state machine, FSM). Особливості побудови і використання моделі FSM надано у Розділі 4.

У наступних підрозділах етапи 1-6, що складаються з отримання наборів, попередньої підготовки даних, вилучення ознак, навчання класифікатора та розпізнавання жестів розглянуто більш докладно.

2.2.1 Отримання набору даних

Для вирішення завдання розпізнавання жестів в дисертаційній роботі створено базу даних відео жестів, яка містить три різних набори даних – два нових набори, записаних і анотованих автором і один вільний набір даних аргентинської мови жестів LSA64 [18], як показано на рис. 2.1.

Набір 1: статичні жести з використанням знакової мови жестів. Набір складається з 30 жестів середньою тривалістю 0.025 сек. Відео жестів отримано зі звичайної веб-камери, що знімала кожен окремий жест з періодичністю 3 мс. Жести виконувалися правою рукою, долоня звернена до камери, приблизно вертикально. Зйомка руки велася з різних кутів огляду. Жести виконували 5 осіб перед веб-камерою. Кожне вхідне зображення містить одну руку. У створеному наборі руки мали колір шкіри європейця.

В результаті попередніх експериментів було виявлено, що бічне світло створює занадто багато контрасту під час бінаризації, в результаті чого частина даних на руці може бути втрачена. Тому в подальшому було вибрано електричне розсіяне світло з регулюванням положення камери для найменшого контрасту між світлими і темними ділянками шкіри. Додатково передбачалося, щоб фон на якому виконувалися жести був менш складним і однорідним.

Для дослідження статичних жестів було використано 12000 зображень, з яких 8400 були випадковим чином взяті для навчального набору, 2400 для валідаційного і 1200 для тестового наборів.

Набір 2: динамічні жести рук для безконтактного перегляду медичних зображень в ОЗ. Набір складається з 10 жестів середньою тривалістю 0.649 сек. Розгорнутий аналіз набору 2 надано у Розділі 4.

Набір 3: знаходиться у вільному доступі [17] і складається з 30 жестів, кожен з яких має 50 відеопотоків. Для обробки цього набору кожен потік був розділений на послідовність кадрів у градаціях сірого. В результаті було створено 90000 зображень, що представляють послідовність динамічних жестів, 3000 для кожної категорії.

Додаткова інформація про особливості використання кожного з наборів, надана у відповідних підрозділах.

2.2.2 Збільшення даних

На цьому етапі до отриманого набору даних застосовується операція збільшення даних (ЗД) (data augmentation). Додаткові зображення генеруються

шляхом маніпуляцій з вже наявними зображеннями, і охоплюють операції масштабування, зрушення, деформацію і повороти під різними кутами тощо. Крім збільшення набору даних для навчання, це робить мережу стійкою до спотворень у вхідних даних на стадії розпізнавання, а також додатково допомагає боротися з перенавчанням [21]. ЗД відноситься до будь-якого методу, який штучно нарощує оригінальний набір із збереженням міток, що зберігає трансформації, і може бути представлений як відображення [19]:

$$\varphi: S \rightarrow T,$$

де S - оригінальний навчальний набір, а T - доповнений набір S.

Тоді, штучно збільшений навчальний набір може бути представлений як:

$$S' = S \cup T$$

де S' містить оригінальний навчальний набір та відповідні перетворення, визначені φ.

При цьому вважається, що якщо зображення x є елементом класу y, тоді $\varphi(x)$ також є елементом класу y. Оскільки існує нескінченний набір відображень $\varphi(x)$, які задовольняють обмеженням збереження міток, у дисертаційній роботі було використано наступні методи ЗД:

Перевертання дзеркально відображає зображення по вертикальній осі. Це одна з найбільш часто використовуваних схем збільшення [11], оскільки є обчислювально ефективною та легко реалізованою оскільки вимагає лише обертання рядків матриць зображень.

Схема *обертання* обертає зображення навколо центру зображення шляхом відображення кожного пікселя зображення (x, y) у (x', y') із наступним перетворенням [3]:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

Експериментально доведено [19], що для створення нових варіантів зразків достатньо встановити кут обертання θ в межах $-30^{\circ} \leq \theta \leq 30^{\circ}$.

Обрізання - ще одна схема збільшення, запропонована у [11]. Використовується та саму процедуру, що і у відповідній роботі, яка полягала у вилученні 224 × 224 посівів з чотирьох кутів та центру зображення 256 × 256.

Фотометричні методи передбачають трансформації, які змінюють канали RGB, шляхом зсуву значень (r, g, b) кожного пікселя до нових значень (r′, g′, b′) пікселів відповідно до заздалегідь визначених евристик. Таке перетворення дозволяє змінювати освітлення та колір зображення, залишаючи незмінним його геометрію. Джітер кольорів - це метод, який використовує випадкові маніпуляції кольорами [23].

Отриманий та збільшений набір даних є вихідними даними для етапу підготовки та обробки даних. Результати ЗД для випадкового кадру надано у Додатку В. Так, наприклад, дані з вихідного набору 1 генерувалися за допомогою афінних перетворень обертання, зрушення і зміну масштабу вихідних зображень з використанням функції Keras ImageDataGenerator [9].

ImageDataGenerator приймає вихідні дані, випадково перетворює їх і повертає лише нові, перетворені дані, тобто без оригінального зображення. Обертання проводилося в діапазоні від 0 до 20 градусів, горизонтальний і вертикальний зсув до 20 fraction по ширині і висоті. Оригінальне і згенеровані зображення надано на рис. 2.2.

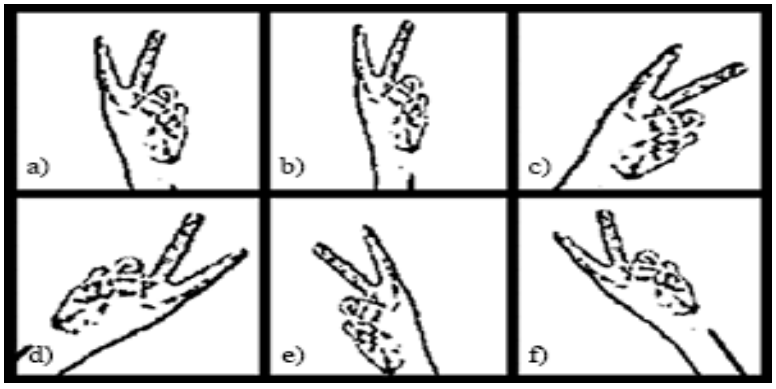


Рисунок 2.2 – Оригінальне зображення (a) і згенеровані зображення (b, c, d, e, f) для Набору 1.

2.2.3 Підготовка та попередня обробка даних

Ґрунтуючись на тому, що операції передобробки, включаючи перетворення зображення, обрізання, перевертання, регулювання кольору та обертання зображення, можуть покращити узагальнення в різних випадках [2], попередня обробка даних для поточного завдання містить очищення даних, зміну колірної моделі, сегментацію та виділення контуру (розпізнавання країв) (рис.2.3). Приклад початкового зображення з набору 1 до обробки представлений на рис. 2.3 а).

Рисунок 2.3 - Підготовка зображення для CNN: (а) початкове зображення, (б) сегментація за кольором шкіри, (с) виділення контурів за Canny

Очищення даних. Перед початком процесу сегментації, необхідно очистити зображення від цифрового шуму і знизити зайву деталізацію. Для цього, до кожного зображення з набору даних було застосовано фільтр розмиття за Гаусом (Gaussian Blur) [14].

$$G(x,y)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}}\ , \tag{2.14}$$

де x – відстань по осі абсцис, y - відстань по осі ординат, σ - стандартне відхилення розподілу Гауса. Взаємний вплив пікселів визначається як обернено пропорційне квадрату відстані між ними. Ступінь розмиття залежить від параметра стандартного відхилення.

Зміна колірної моделі. Перетворення кольорового простору RGB можливе у кольоровий простір HSV або YCbCr. У просторі HSV хроматична інформація зберігається в окремому каналі, що дозволяє легше орієнтуватися при сегментації на колір шкіри. Також при використанні HSV знижуються вимоги з мінливості освітлення.

Вищезазначений процес можна математично описати наступним чином: Нехай загальна кількість пікселів дорівнює N, а H, S та V представляють відтінок (H – hue), насиченість (S – saturation) та значення яскравості (V – value або B - brightness) відповідно. Якщо прийняти, що MAX дорівнює максимальному значенню, кожного з каналів RGB, а MIN – мінімальному, то перетворення в формат HSV здійснюється наступним чином:

$$H = \begin{cases} 60 \times \dfrac{G-B}{MAX-MIN} + 0, if\ MAX = R, G \geq B \\ 60 \times \dfrac{G-B}{MAX-MIN} + 360, if\ MAX = R, G < B \\ 60 \times \dfrac{B-R}{MAX-MIN} + 120, if\ MAX = G \\ 60 \times \dfrac{R-G}{MAX-MIN} + 240, if\ MAX = B \end{cases}$$

$$S = \begin{cases} 0, & \text{if } MAX = 0 \\ 1 - \dfrac{MIN}{MAX}, & \text{otherwise} \end{cases}$$

$$V = MAX$$

Система повинна вимагати трьох обмежень для досягнення кольорового порогу параметрів H, S та V: (1) по H обрати область кольору шкіри, (2) насиченість S має бути не білою, тобто більше певного порогу $\mathcal{T}s$, і (3) область кольору шкіри повинна бути більш яскравою V у порівнянні з будь-яким іншим предметом, який має подібний колір із шкірою, тобто також більше відповідного порогу $\mathcal{T}v$. Отже, для кожного пікселя зображення, обмеження по кольоровому простору HSV можна узагальнити наступним чином:

$$(H,S,V) = \begin{cases} -10 < H < 10, \\ S > \mathcal{T}_S, \\ V > \mathcal{T}_S. \end{cases}$$

Альтернативним варіантом перетворення колірної моделі RGB є кольоровий простір YCbCr. Формат YCbCr може бути використаний для двох задач: (1) для масштабування значень, які менше 255, що дозволяє зміщувати компоненти кольоровості і забезпечити неупереджене подання; (2) для максимізації динамічного діапазону при роботі з 8-бітними вхідними зображеннями (від 0 до 255):

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.500 \\ 0.500 & -0.419 & -0.081 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 0 \\ 128 \\ 128 \end{bmatrix}$$

Значення кольору шкіри людини в кольоровому просторі YCbCr визначено в межах таких обмежень [27]:

$$\begin{cases} 75 \leq Cb \leq 140, \\ 130 \leq Cr \leq 180. \end{cases}$$

Пікселі, що знаходяться у визначеному діапазоні, розглядаються як шкіра і передаються на наступний етап для сегментації.

Процес сегментації. Сегментація полягає у відокремленні руки від фону що досягається шляхом використання порогу. Припустимо, що вхідне зображення дорівнює $f(x, y)$, вихідне значення $f'(x, y)$, поріг - $\mathcal{T}$, процес сегментації з використанням порогового значення можна описати за такою формулою:

$$f'(x,y) = \begin{cases} \alpha_1 & f(x,y) < \mathcal{T}, \\ \alpha_2 & f(x,y) \geq \mathcal{T}. \end{cases}$$

Процес сегментації на основі порогу, дає два значення зображення. Особливостями зображення є, як правило, колір, шкала сірого або інша

інформація про зображення. Припустимо, що у загальному випадку, поріг може бути описаний таким чином:

$$\mathcal{T} = \mathcal{T}(Poz(x,y), Gr(x,y), X(x,y)).$$

При $\mathcal{T} = \mathcal{T}\big(Gr(x,y)\big)$, вважається, що поріг пов'язаний лише зі шкалою сірого. При $\mathcal{T} = \mathcal{T}(Gr(x,y), X(x,y))$, поріг визначається шкалою сірого та іншими характеристиками. Повний запис $\mathcal{T} = \mathcal{T}(Poz(x,y), Gr(x,y), X(x,y))$ означає, що поріг визначатиметься положенням (*Poz*), шкалою сірого (*Gr*) та іншими характеристиками (*X*). Оскільки введення жестів рукою є гнучким і складним, не існує універсального методу, який можна застосувати до всього процесу сегментації жестів. За умов використання CNN, належний поріг для сегментації може бути визначений за допомогою тренувального набору. Вибір порогу може мінімізувати шум і максимальну швидкість виявлення кольору шкіри.

Сегментація в запропонованій системі розпізнавання жестів здійснюється за алгоритмом Otsu [15], який є непараметричним, некерованим методом автоматичного вибору порогу.

Алгоритм трактує сегментацію зображення як проблему бінарної класифікації, в якій два класи (рука і фон) генеруються із набору пікселів у зображенні шкали сірого. Існує L рівнів у зображенні в масштабі променя (0-255). Використовуючи поріг ***T*** для зображення з *L* рівнями сірого, зображення сегментується за двома класами $\Omega_0 = (1,2\ldots. k)$ та $\Omega_0 = (k+1, k+2\ldots., L)$.

Оптимальний поріг k^* визначається як значення k, яке максимізує відношення дисперсії між класами σ_B^2B до загальної дисперсії σ_k^2. після знаходження порогового значення k пікселям рук було призначено «1», а фоновим пікселям - «0».

В результаті сегментації відокремлюється задній фон, і подальша робота ведеться тільки з рукою. Зображення перетворюється в одноканальне у відтінках

сірого для зменшення обчислювальних витрат. Приклад сегментації за кольором шкіри представлений на рис. 2.2 б).

Розпізнавання контурів. Останньою процедурою підготовки було розпізнавання контурів. Розпізнавання контурів є важливим моментом в розпізнаванні жестів, визначаючи границю між об'єктами або між об'єктом і фоном. До виділеного елементу застосовується детектор контурів Canny [7]. Робота детектора Canny складається з наступних кроків:

1. Згладжування зображення шляхом застосування до нього розглянутого вище фільтра Гаусова розмиття.

2. Взяття градієнта зображення, після чого на максимальних значеннях позначаються межі. Для цього використовується оператор Собеля [20], що обчислює значення градієнта яскравості. Оператор Собеля використовує два квадратних ядра згортки, які оцінюють градієнт в горизонтальному і вертикальному напрямках. Після проходження згортки напрямок градієнта обчислюється як:

$$\Theta = \arctan\left(\frac{G_y}{G_x}\right), \tag{2.15}$$

де G_y і G_x – значення для першої похідної відповідно в горизонтальному і вертикальному положеннях.

3. Видалення немаксимумів. Краями визнаються пікселі, в яких досягається локальний максимум градієнта в напрямку вектору градієнта. Значення кожного пікселя невизнаного максимумом встановлюється в нуль. В результаті виходить тонка лінія контуру.

4. Подвійна порогова фільтрація. Для оцінки того, чи дійсно має місце край в конкретній точці зображення, використовуються два порога. Якщо значення пікселя відноситься вище порога, то край визнається достовірним. Інакше відкидається. Проміжним пікселям присвоюється середнє значення.

5. Гістерзіс, тобто зв'язування країв в контури. Пікселі вище порогового значення T_1 є крайовим пікселем. Пікселі, які межують з крайовим пікселем, і

при цьому мають значення вище, ніж T_2 також відносяться до групи крайових пікселів.

В результаті виконання даного етапу отримується набір зображень, готових для завантаження до нейронної мережі. Приклад результуючого зображення, підготовленого до розпізнавання, представлено на рис. 2.2 с.

2.3 Модель обробки потокових даних для розпізнавання статичних жестів із доповненими даними

Для вирішення задачі (2.1), в роботі використовується класифікатор на базі згорткової нейронної мережі (CNN), яка дозволяє розв’язати два завдання: (1) визначення ознак об'єктів, (2) визначення ймовірності приналежності об'єкта (зображення X_t) до того чи іншого класу жестів S_i.

2.3.1 Архітектура мережі для статичного розпізнавання жестів

На відміну від інших методів машинного навчання (МН), згорткова нейронна мережа не вимагає ручної розробки набору ознак. Ознаки витягуються мережею самостійно в згорткових шарах. Модель пропонованої згорткової нейронної мережі надано на рис. 2.4.

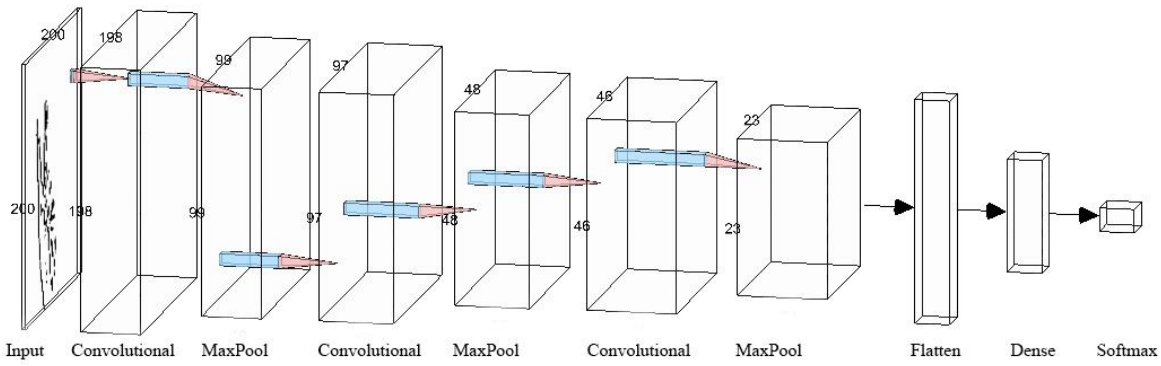


Рисунок 2.4 – Архітектура згорткової нейронної мережі для розпізнавання статичних жестів

Розроблена мережа має три згорткових шари з функцією активації ReLU, за кожним з яких розташовується субдіскретізаційний шар максимізаційного агрегування max pooling. Умовно, створену CNN можна розділити на три частини: а) три блоки згорткових і субдіскретізаційних шарів, які чергуються один з одним і формують карти ознак; б) три повноз'єднаних шари, що перетворюють вхідні дані у вектор ознак і на виході класифікують зображення. Згорткові шари, формують 16, 32 і 64 карти ознак, застосовуючи ядра згортки розміром 3×3. Карти ознак проходять через субдіскретізаційні шари з максимізаційним агрегуванням 2×2, кожен раз зменшуючи розмірність даних вдвічі. Дана мережа містить 10 шарів. До трьох згорткових шарів застосовуються відповідно 16, 32 і 64 згорткових ядра без доповненням зображення нульовими пікселями, тобто padding шара дорівнювався valid. Функцією активації є Rectified Linear Unit (ReLU), яка вносить нелінійність у модель.

За кожним згортковим шаром іде шар максимізаційного агрегування розміром 2×2 на кожен крок. Після чого, дані в згладжувальному шарі (flatten) перетворюються з 2D-представлення в одновимірний вектор, і в кінці проходять через два повноз'єднаних шари.

Другий повноз'єднаний 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 функції втрат категорійної крос-ентропії.

2.3.2 Формування ознак

У згорткових шарах ядра згортки застосовують лінійне перетворення до кожного пікселя зображення, виявляючи характерні ознаки, властиві даному класу зображень. Ядра являють собою матриці ваг непарних розмірів, чиї значення встановлені так, щоб можна було виявити будь-які просторові або

кольорові ознаки. На виході маємо карти ознак, чия кількість дорівнює кількості застосованих згорткових ядер [16]. Згортку можна уявити, як (2.16).

$$(I * K)_{ij} = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} \omega_{ab} y^{l-1}_{(i+a)(j+b)}, \tag{2.16}$$

де ω – ядро розміру m×m, y – входи з попереднього шару, f – функція активації нейронів.

Розмір вихідних даних із шару згортки обчислюється за формулою (2.17).

$$n_{out} = \left\lfloor \frac{n_{in} + 2p - k}{s} \right\rfloor, \tag{2.17}$$

де n_{in} – розмір вхідних даних з попереднього шару, p – відступ, k – розмір ядра, s – розмір кроку, з яким ядро зміщується по зображенню.

Ядра згортки формують карти ознак, що дорівнюють їх кількісті. У даному випадку були використані три згорткових шара, які давали на виході 16, 32 і 64 карти ознак відповідно. Використання трьох згорткових шарів представляється найбільш оптимальним щодо якості навчання до малої кількості параметрів. Кількість параметрів для згорткового слою визначається як (2.18).

$$p = d * k * w + b, \tag{2.18}$$

де d – глибина вхідних даних (кількість каналів), k – ядро, w – ваги, b – bias. Наприклад, для першого згорткового шару це буде виглядати як $1*16*9+16$, тобто 160 параметрів. Для другого шару $16*32*9+32$, тобто 4640 параметрів.

Результат згортки необхідно пропустити через нелінійну функцію активації. Нелінійність можна записати у вигляді виразу (2.19).

$$y^l_{ij} = f(x^l_{ij}). \tag{2.19}$$

У цьому випадку функцією активації виступала Rectified Linear Unit ReLU [1].

$$f(x_i) = \begin{cases} x_i, & x_i > 0 \\ a_i x_i, & x_i \le 0 \end{cases}, \tag{2.20}$$

де x_i – вхідні дані і-го каналу, a_i – коефіцієнт, який контролює нахил на негативних значеннях. ReLU значно перевершує інші функції в стійкості до згасання градієнта, а швидкість навчання згідно [11] в порівнянні з гіперболічним тангенсом швидше в шість разів.

Сформовані на кожному шарі карти ознак подаються на шар субдіскретізації max-pooling.

Шар max-pooling з фільтром 2x2 слідує за кожним шаром згортки, підсилює виявлені на попередньому шарі ознаки і зменшує розмірність вхідних даних і, як наслідок, кількість параметрів. Максимізаційне агрегування обчислюється як (2.21).

$$y^{l+1} = \max_{0 \le i < H, 0 \le j < W} x^{l}_{i^{l+1} \times H + i, j^{l+1} \times j, d}, \tag{2.21}$$

де H и W – висота і ширина фільтра субдіскретізації, x – вхідні дані.

Розмір вихідних даних із шару max-pooling визначається як (2.22).

$$n_{out} = \left\lfloor \frac{n_{in} - k}{s} \right\rfloor + 1. \tag{2.22}$$

Максимізаційне агрегування не має ніяких параметрів для навчання.

З метою запобігання перенавчання, при якому навчена модель занадто точно відповідає заданому набору даних, втрачаючи здатність до узагальнення, в розробленій нейромережі застосовується виключення dropout [5]. Dropout дозволяє на кожному етапі навчання виключати в певному шарі випадкові

нейрони із заданою наперед ймовірністю. Виключений нейрон повертає значення 0. Імовірність того, що будь-якої нейрон на поточну епоху залишиться в мережі становить q = 1 – p.

Цей прийом дозволяє нейромережі не просто запам'ятовувати правильні відповіді, а надає гнучкість в розпізнаванні зразків, що не включені до неї ні в один з навчальних датасетів [25]. Окремі вузли виключаються з мережі із установленою ймовірністю 0.25, тобто на кожній ітерації 25 відсотків випадкових нейронів будуть виключатися з роботи мережі.

2.3.3 Класифікація даних

Класифікуюча частина мережі складається зі згладжувального шару flatten, кожен вузол якого відповідає одному значенню з вектора ознак, і двох повноз'єднаних шарів dense. Останній шар є вихідним і реалізує функцію softmax. Класифікуюча частина являє собою багатошаровий персептрон і призначена для навчання класифікатора жестів рук за виявленими ознаками. Дані з ознаками з останнього шару субдіскретізації надходять на згладжувальний шар, перетворюючись тим самим в одномірний вектор. Оскільки на вхід згладжувального шару надходить 64 карти ознак розміром 13×13, то перший шар повноз'єднаної частини мережі складається з 10816 нейронів.

Наступний шар є повноз'єднаним і містить 256 ваг, що цілком достатньо для вирішення поставленого завдання, не вимагаючи багато ресурсів. В результаті кількість параметрів на цьому шарі становить 10816×256 + 256 = 2769152 параметрів.

Обчислення значень нейронів для повноз'єднаного шару відбувається за формулою (2.23).

$$x_i^l = \sum_{k=0}^{m} w_{ki}^l y_k^{l-1} + b_i^l, \qquad (2.23)$$

де w_{ki}^{l} – вага від k-го нейрона шару *l*-1 до i-того нейрону поточного шару *l*, *b* – зміщення поточного шару, y – вхідні дані з попереднього шару.

Останній шар мережі з кількістю виходів дорівнює кількості розпізнаваних категорій, реалізує функцію активації softmax.

Softmax привласнює значення, представлене позитивним дійсним числом кожного класу, виражаючи ймовірність приналежності наступним чином (2.24).

$$S_i = P(y = i \mid a), \tag{2.24}$$

де y – вихідний клас пронумерований як 1..n, a – вектор розмірності n.

Сума всіх вихідних сигналів, що виражають ймовірність приналежності, дорівнює одиниці. Значення вихідного сигналу *i*-го нейрона відповідає ймовірності того, що правильна відповідь є *i*:

$$y_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{2.25}$$

де n – кількість класів z_i це the i-ий єлемент z, а $y = [y_1, y_2, \ldots, y_c]^T$ – вихід шара классификатора softmax. Значення S_i завжди позитивно і знаходиться в діапазоні (0, 1).

2.3.4 Навчання мережі

Навчання мережі відбувається на основі алгоритму зворотного поширення помилки. Зворотне поширення можна розглядати як диференціювання складної функції з пошуком похідних втрат по змінним. У якості функції втрат була використана категорійна перехресна ентропія – розрахована логарифмічна втрата для декількох представлених класів. Перехресна ентропія між розподілами *p* і *q* визначається наступним чином (2.26).

$$H(p,q) = H(p) + D_{KL}(p \parallel q), \tag{2.26}$$

де $H(p)$ - ентропія p, $D_{KL}(p||q)$ - розбіжність Кулбака-Лейблера [12] для q з p (відносна ентропія p до q).

Якщо прогнозовані значення моделі дорівнюють q, тоді як справжні значення дорівнюють p, то категорійна перехресна ентропія буде виглядати наступним чином (2.27).

$$H(y,\hat{y}) = -\sum_i y_i \log \hat{y}_i = -y \log \hat{y} - (1-y)\log(1-\hat{y}), \tag{2.27}$$

де у – бажаний результат, $\hat{y}$ – фактичний результат, між якими необхідно мінімізувати перехресну ентропію. Для навчальної вибірки розміром m мінімізація перехресної ентропії між передбаченим і реальним значенням буде виглядати як (2.28).

$$E(w) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)], \tag{2.28}$$

де m – кількість класів. У разі, якщо вихідний сигнал близький до очікуваного, значення функції втрат близько до нуля $\lim_{\hat{y}\to y} E = 0$.

Залежно від результатів обчислення функції втрат на кожній ітерації параметри змінюються через (2.29).

$$w_{ij}^{l} = w_{ij}^{l} - \alpha \frac{\partial}{\partial W_{ij}^{l}} E(W,b), \tag{2.29}$$

де α – швидкість навчання.

Зміна значень ваг буде здійснюватися через градієнтний спуск (2.30).

$$\Delta w_{ij} = -\alpha \frac{\partial E}{\partial w_{ij}}. \quad (2.30)$$

Таким чином, завдання навчання класифікатора S полягає в мінімізації функції втрат в просторі ваг (2.31).

$$\min_{W} E(S(X,W)M), \quad (2.31)$$

де $\boldsymbol{X}$ – простір ознак, W – ваги мережі, M – множина класів. Виходячи з цього, необхідно обчислити градієнт функції втрат (2.32).

$$\nabla E(S(X,W)M). \quad (2.32)$$

При виборі оптимізатора були випробувані два адаптивних методи – Adadelta [26] і Adam [10]. Власне, Adadelta є розширенням іншого оптимізатора – AdaGrad [4], у якого виникала проблема зі зменшенням швидкості навчання. Проблема виникала внаслідок накопичення суми квадратів градієнтів, що видно із самої формули AdaGrad (2.33).

$$w_{N+1} = w_N - \frac{\alpha}{\sqrt{G_N + \varepsilon}} g_N, \quad (2.33)$$

де w_N – значення параметра w на кроці N, G_N – діагональна матриця, що містить суму квадратів оновлень часткових похідних параметра w від початку роботи до N, ε – згладжуючий параметр для уникнення поділу на нуль.

У Adadelta замість повної суми оновлень використовується усереднений квадрат градієнта, для чого використовується експоненціально затухаюче рухоме середнє. Експоненціальне середнє вираховується як (2.34).

$$E[g^2]_N = \gamma E[g^2]_{N-1} + (1-\gamma) g_N^2, \quad (2.34)$$

де γ – коефіцієнт загасання в часі. Оскільки середнє квадратів градієнтів виглядає як (2.35)

$$RMS = \sqrt{E[g^2]_N + \varepsilon}\,, \tag{2.35}$$

то правило поновлення ваг для Adadelta приймає вигляд (2.36)

$$w_{N+1} = w_N - \frac{RMS[\Delta w]_{N-1}}{RMS[g]_N} g_N\,. \tag{2.36}$$

Другим випробуваним оптимізатором градієнтного спуску був метод адаптивної інерції (Adaptive momentum estimation, Adam) [10]. У багатьох практичних додатках Adam пропонується як алгоритм оптимізації за замовчуванням для глибокого навчання [8]. Він має більше гіперпараметрів порівняно з іншими оптимізаторами, але добре працює з невеликим налаштуванням гіперпараметрів [25]. Adam обчислює адаптивні швидкості навчання, виходячи з оцінок першого і другого моментів градієнтів. Оцінка першого моменту обчислюється виходячи з отриманих раніше значень часткових похідних, як рухома середня градієнтів. Оцінка другого моменту обчислюється на основі квадратів градієнтів значень останніх величин для ваги (2.37)-(2.38).

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t\,, \tag{2.37}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2\,, \tag{2.38}$$

де β_1 – коефіцієнт експоненційної швидкості затухання для першого моменту, β_2 – коефіцієнт експоненціального спаду для оцінки другого моменту. У зв'язку з тим, що оцінки моментів инициализируются нулями, а значення коефіцієнтів β_1 і β_2 рекомендуються авторами близькими до одиниці (0.9 і 0.999 відповідно),

зберігається зрушення до нуля. Для запобігання цьому для зміни параметрів слід використовувати (2.39) – (2.40) [6].

$$\hat{m}_p = \frac{m_p}{1-\beta_1^p}, \tag{2.39}$$

$$\hat{v}_p = \frac{v_p}{1-\beta_2^p}. \tag{2.40}$$

Перерахунок параметрів здійснюється за формулою (2.41).

$$w_p = w_{p-1} - \frac{\alpha}{\sqrt{v_p + \varepsilon}} m_p, \tag{2.41}$$

де $\varepsilon=10^{-8}$ і вводиться для запобігання можливого поділу на нуль. У нашому випадку швидкість навчання $\alpha = 0.002$. Значне збільшення швидкості знижувало ефективність навчання мережі і збільшувало кількість невірних відповідей. Зменшення швидкості призводило до довгого навчання мережі.

Проведені експерименти показали досить високу ефективність Adadelta, однак, при порівнянні з Adam навчання тривало більш ніж в два рази довше.

Маючи значення помилок, далі обчислювалася часткова похідна цільової функції E по відношенню до кожного виходу нейрона. Щоб обчислити похідні втрат по змінним у вкладеному рівнянні, застосовується ланцюгове правило (2.42).

$$\frac{\partial E}{\partial x_{ij}^l} = \frac{\partial E}{\partial y_{ij}^l}\frac{\partial y_{ij}^l}{\partial x_{ij}^l} = \frac{\partial E}{\partial y_{ij}^l}\frac{\partial}{\partial x_{ij}^l}(f(x_{ij}^l)) = \frac{\partial y_{ij}^l}{\partial x_{ij}^l} f(x_{ij}^l). \tag{2.42}$$

Для згорткового шару процедура зворотного поширення помилки виглядає як (2.43).

$$\frac{\partial E}{\partial y_{ij}^{l-1}} = \sum_{a=0}^{m-1}\sum_{b=0}^{m-1} \frac{\partial E}{\partial x_{(i-1)(j-b)}^{l}} \frac{\partial_{(i-1)(j-b)}^{l}}{\partial y_{ij}^{l-1}} = \sum_{a=0}^{m-1}\sum_{b=0}^{m-1} \frac{\partial E}{\partial x_{(i-1)(j-b)}^{l}} w_{a,b}. \quad (2.43)$$

В шарі максимізаційного агрегування помилка від попереднього шару проходить через єдине максимальне значення. Оскільки цей шар не використовується для навчання мережі, то і помилка через нього проходить без змін.

Для шару softmax необхідно продиференціювати функцію втрат J по відношенню до z_i (2.44).

$$\frac{\partial J}{\partial z_i} = -\sum_k y_k \frac{\partial \log \hat{y}_k}{\partial z_i} = -\sum_k y_k \frac{1}{\hat{y}_k}\frac{\partial \hat{y}_k}{\partial z_i} = -y_i(1-\hat{y}_i) - \sum_{k\neq i} y_k \frac{1}{\hat{y}_k}(-\hat{y}_k\hat{y}_i)$$
$$= -y_i(1-\hat{y}_i) + \sum_{k\neq i} y_k(\hat{y}_i) = -y_i + y_i\hat{y}_i + \sum_{k\neq i} y_k(\hat{y}_i) = \hat{y}_i(\sum_k y_k) - y_i = \hat{y}_i - y_i \quad . (2.44)$$

Результати тестування мережі наведено у Розділі 5.

Враховуючи той факт, що дослідження проводилася на двоядерному i3-7100 процесорі з використанням простої веб-камери отримано високий ступінь вірних прогнозів.

Серед недоліків моделі варто відзначити, що запропонована методологія підходить лише до випадків жестів, присутніх на статичних зображеннях, без виявлення та відстеження рук або випадків оклюзії рук. У наступному підрозділі для розпізнавання складніших жестів до CNN доданий рекурентний блок.

2.4 Модель обробки потокових даних для розпізнавання динамічних жестів

Виходячи з постановки, задача динамічного розпізнавання полягає не лише в тому, щоб знайти цільове зображення у будь який момент часу та

відокремити його від фону, а також в аналізі динамічних функцій простору-часу, та відстежуванні початку і кінця жесту в потоці наступних кадрів.

З цією метою, модель розпізнавання на базі CNN доповнюється мережею довгострокової пам’яті (LSTM). Перший компонент системи, CNN, призначений для вилучення ознак із глибинного зображення суб'єкта, який виконує жест рукою. Другий компонент дозволяє виконувати вилучення часових ознак і, тим самим проводити динамічне розпізнавання жестів яке залежить від упорядкованої послідовності зображень.

2.4.1 Архітектура мережі для динамічного розпізнавання жестів

Загальна архітектура мережі CNN-LSTM наведена на рис. 2.5. Частина мережі, що відповідає за виявлення просторових ознак (spatial features) містить три згорткові шари (CNN), за якими йдуть шари агрегування (MaxPool), згладжувальний (flatten), шари виключення (Dropout), і один вихідний повноз'єднаний шар (FC).

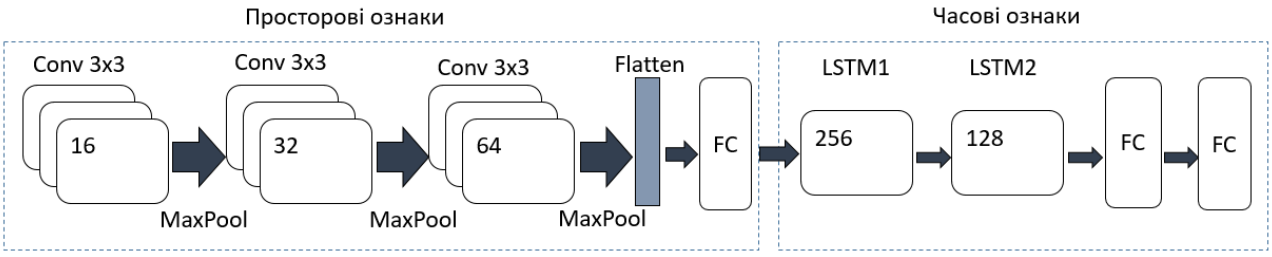


Рисунок 2.5 - Архітектура нейронної мережі CNN-LSTM для динамічного розпізнавання жестів

На виході з CNN-частини мережі в згладжувальному шарі дані перетворюються з 2D-зображення в одновимірну векторну форму, і в кінці вони проходять через повноз'єднаний щільний шар.

З метою розпізнавання часових послідовностей у відеопотоках, до CNN додаються шари з LSTM [22].

Вхідними даними є просторові ознаки, виявлені на попередніх згорткових шарах, сплощених в одновимірний вектор. Шари LSTM з 256-а та 128-ю вентильнимі вузлами визначають часові патерни на 60 часових кроках.

2.4.2 Навчання мережі

Для вирішення задачі розпізнавання жестів застосовувалась CNN, навчена через функцію категоріальної перехресної ентропії (categorical cross-entropy, CCE). На першому етапі було використано власний набір зображень жестів рук, далі, щоб уникнути надмірного спрощення, оскільки отримані зображення були досить простими, у було проведено перевірку моделі на загальнодоступному наборі даних.

Процес навчання починається зі зчитування всіх зображень з бази даних і передбачає виконання наступних етапів (рис. 2.6):

- побудова бази даних розмічених даних жестів рук;
- попередня обробка даних;
- збільшення даних;
- виявлення просторових ознак у згорткових шарах;
- виявлення тимчасових ознак у шарах LSTM;
- розпізнавання жесту руки.

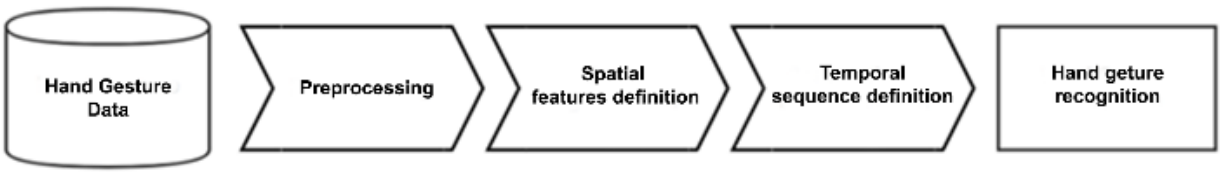


Рисунок 2.6 - Схема розпізнавання динамічного жесту

На другому етапі виконується попередня обробка даних, зображення перетворюються у відтінки сірого, сегментуються та спрощуються, як це описано в п. 2.2. Цей процес дозволяє усунути незначущі деталі та зосередитись

лише на інформативних ознаках. Після попередньої обробки вирішується проблема ідентифікації візуальних ознак у кожному кадрі.

Класифікація проводиться у згортковому та агрегувальному шарах. Виявлені ознаки з останнього шару max-pooling згладжуються в одномірний масив і подаються до вхідних шарів LSTM, які визначають їх часові залежності.

Заключний етап складається з двох операцій: обчислення функції втрат і оптимізації. Функція втрат показує різницю між прогнозованою та спостережуваною величинами. У свою чергу, завдання оптимізатора полягає в оновлені значення ваг для зменшення вартості функції збитків.

У якості функції втрат застосовувалась категоріальна перехресна ентропія. Рівнянням для категоріальної перехресної ентропії є (2.45)

$$L(y,p) = -\frac{1}{O}\sum_{o=1}^{O}\left(y_o \log(p_o) + (1-y_o)\log(1-p_o)\right) = -\frac{1}{O}\sum_{o=1}^{O}\sum_{c=1}^{C} y_{oc}\log(p_{oc}). \quad (2.45)$$

Результати тестування мережі наведено у Розділі 5.

Висновки до другого розділу

1. У розділі запропоновано нову методологію розробки і спільного використання візуальних методів розпізнавання жестів рук, яка дозволяє створювати і досліджувати моделі глибокого навчання для розпізнавання статичних та динамічних жестів, здатних працювати в режимі реального часу і забезпечує розуміння способів їх налаштування для різних інтерфейсів управління жестами та потенційних застосувань в системах ЛМВ.

2. Надано приклад використання методології для розробки моделей розпізнавання статичних та динамічних жестів рук, побудованих на основі CNN та CNN+LSTM архітектур для використання в режимі реального часу в інтерактивних людино-машинних інтерфейсах. Представлено математичний апарат для реалізації розпізнавання жестів рук, та основні процедури витягу

часово-просторових ознак в даних.

3. Завдання класифікації жесту сформульовано у вигляді задачі пошуку приналежності зображення до певного класу жестів, для якого розподіл ймовірності має максимальне значення

4. Удосконалено модель статичного розпізнавання жестів руками за допомогою CNN. До набору даних було застосовано технологію збільшення даних, таке як повторне масштабування, масштабування, зсув, обертання, зміщення ширини та висоти і операцію контурування. Модель була навчена на 8000 зображеннях та протестована на 1600 зображеннях, які були розділені на 10 класів. Модель із доповненими даними досягла точності 97,12%, що майже на 4% перевищує модель без доповнень (92,87%).

5. 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.

6. При виборі оптимізатора було випробувано два адаптивних методи – Adadelta і Adam. Проведені експерименти підтвердили високу ефективність Adadelta, однак, при порівнянні з Adam він показав більш ніж в два рази довше навчання мережі.

7. Для розпізнавання динамічних жестів до CNN додано рекурентний блок. В результаті експериментів, на тестовій підмножині розміром в 1535 зображень була досягнута точність розпізнавання 94%. В ході тестування з розпізнаванням жестів на веб-камеру в режимі реального часу була отримана висока ступінь вірних відповідей мережі, в тому числі, на людях, які не брали участі у створенні бази даних зображень.

8. Запропонована система гнучка, слідуючи підходу, запропонованому в п. 2.2 її можна розширити, додавши нові жести, або зменшити, видаливши деякі жести.

Результати розділу опубліковано в роботах автора [4, 5, 7, 8, 10] (Додаток А).

Література до другого розділу

1. Agarap A.F. (2018). Deep learning using rectified linear units (ReLu). In: *arXiv preprint* arXiv:1803.08375.

2. DeVries T., Taylor G.W. (2017) Improved regularization of convolutional neural networks with cutout. In: *arXiv preprint* arXiv:1708.04552.

3. Dieleman S., Willett K., Dambre J. (2015) Rotation-invariant convolutional neural networks for galaxy morphology prediction. In: *Monthly Notices of the Royal Astronomical Society*, pp. 1441-1459.

4. Duchi J., Hazan E., Singer Y. (2011) Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12, pp. 2121-2159.

5. Hinton G.E., et al. (2012) Improving neural networks by preventing co-adaptation of feature detectors. In: *arXiv preprint* arXiv:1207.0580.

6. Ilievski I., Ahkar T., Feng J., Shoemaker Ch.A. (2017) Efficient Hyperparameter Optimization of Deep Learning Algorithms Using Deterministic RBF Surrogates. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pp. 822-829.

7. Canny J. (1986) A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, vol. 6, pp. 679-698.

8. Karpathy A. et al. (2016) Cs231n convolutional neural networks for visual recognition. *Neural networks*, 1:1.

9. Keras - Image Data Preprocessing. Режим доступу: https://keras.io/api/preprocessing/image/ (16.08.2019).

10. Kingma D.P., Ba J. (2014) Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*.

11. Krizhevsky A., Sutskever I., Hinton G.E. (2017) ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6, pp. 84-90. DOI: https://doi.org/10.1145/3065386.

12. Kullback S., Leibler R.A. (1951) On information and sufficiency. *The annals of mathematical statistics*, vol. 22.1, pp. 79-86.

13. Lupinetti K., Ranieri A., Giannini F., Monti M. (2020) 3D Dynamic Hand Gestures Recognition Using the Leap Motion Sensor and Convolutional Neural Networks. In: *De Paolis L., Bourdot P. (eds) Augmented Reality, Virtual Reality, and Computer Graphics. AVR 2020. Lecture Notes in Computer Science*, vol 12242. Springer, Cham. https://doi.org/10.1007/978-3-030-58465-8_31.

14. Nixon M., Aguado A.S. (2012) Feature extraction and image processing for computer vision. *Academic Press*.

15. Otsu N. (1979) A Threshold Selection Method from Gray-Level Histograms, *IEEE transactions on systems, man, and cybernetics*, vol. smc-9, no. 1.

16. Pinto R.F., Borges C.D.B., Almeida A.M.A., Paula I.C. (2019) Static Hand Gesture Recognition Based on Convolutional Neural Networks, *Journal of Electrical and Computer Engineering*, vol. 2019, Article ID 4167890, 12 p. https://doi.org/10.1155/2019/4167890.

17. Quiroga F. LSA64: A Dataset for Argentinian Sign Language. Режим доступу: http://facundoq.github.io/datasets/lsa64/

18. Ronchetti F., Quiroga F., Estrebou C.A., Lanzarini L.C., Rosete A. (2016). *LSA64: An Argentinian Sign Language Dataset*. Web

19. Taylor L., Nitschke G. (2018) Improving Deep Learning with Generic Data Augmentation, *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, Bangalore, India, 2018, pp. 1542-1547, doi: 10.1109/SSCI.2018.8628742.

20. Sobel I., Feldman G. (1968) A 3x3 isotropic gradient operator for image processing. *A talk at the Stanford Artificial Project*, pp. 271-272.

21. Wang J., Perez L. (2017) The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, p. 11.

22. Wang S., et al. (2016) Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum. *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 851-860.

23. Wu R., Yan S., Shan Y., Dang Q., Sun G. (2015) Deep image: Scaling up image recognition. In: *arXiv preprint* arXiv:1501.02876.

24. Xu P. (2017) A Real-time Hand Gesture Recognition and Human-Computer Interaction System. In: *arXiv preprint* arXiv:1704.07296v1

25. Yu T., Zhu H. (2020) Hyper-Parameter Optimization: A Review of Algorithms and Applications. In: *arXiv preprint* arXiv:2003.05689v1.

26. Zeiler M.D. (2012) ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv*:1212.5701.

27. Zhou W., Lyu C., Jiang X., Li P., Chen H., Liu Y.-H. (2017). Real-time implementation of vision-based unmarked static hand gesture recognition with neural networks based on FPGAs. *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO).* doi:10.1109/robio.2017.8324552.

РОЗДІЛ 3
РОЗРОБЛЕННЯ ТЕХНОЛОГІЇ РОЗПІЗНАВАННЯ ТА ПРОГНОЗУВАННЯ ЖЕСТІВ НА ОСНОВІ МОДЕЛІ ГЕНЕРАЦІЇ ПОСЛІДОВНОСТЕЙ

У розділі продемонстровано удосконалену модель ConvLSTM2D + Conv3D для передбачення жестів рук за допомогою штучної нейронної мережі, використовуючи дані жестів, отриманих з відео. Формалізовано процес прогнозування відеопослідовностей зображень жестів. Розкрито особливості специфікації нейронної мережі, визначено критерії якості прогнозування відеопослідовностей зображень жестів, представлено архітектуру запропонованої мережі.

3.1 Постановка задачі прогнозування жестів

На відміну від розпізнавання жестів, модель прогнозування жестів у реальному часі в цьому дослідженні дозволяє визначити намір руху руки та *передбачати кінцевий жест до кінця руху* руки. Модель прогнозування жестів може бути використана для оцінки жестів людини до фактичної дії, щоб зменшити затримки в інтерактивних системах ЛМВ.

У цій частині ставиться задача навчити модель прогнозувати наступні кадри по набору вже проглянутих протягом виділеного часу. Часовим інтервалом виступає частота кадрів у відеопослідовності довжиною *t*.

Ознаки кадрів з 1 до $t-1$ використовуються для прогнозування ознак на наступному кадрі *t*. Проблема прогнозування формулюється наступним чином [11].

Нехай відома множина кадрів S, представлених сіткою розміром $M \times N$, що складається з M рядків та N стовпців. Всередині кожної клітинки сітки є P вимірювань, які змінюються з часом. X – тензор ознак кадру $X \in \mathbb{R}^{P \times M \times N}$, де $\mathbb{R}$ означає простір спостережуваних ознак. $X_{\times}S$ – просторово-часові послідовності, де $S = \{s_1, s_2, \ldots, s_i\}$ – просторова сітка, та $T = \{t_1, t_2, \ldots, t_i\}$ – тривалість

послідовності навчальних кадрів. C – тензор прогнозування ознак кадру *C*, де *C*(*s*,*t*) – ознаки кадру в сітці $s \in S$ в часі $t \in T$. *X* – список тензорів ознак $X = \{x_1, x_2, ..., x_T\}$, де x_i являє собою тривимірний тензор, що записує відповідний атрибут в кожну сітку s_i на кожному кадрі.

Для набору даних, де D_{train} – тренувальні дані $D_{train} = \{C(S,t), X(S,t_J)\}$ де $t \in T_{train}$, D_{val} – валідаційні дані $D_{val} = \{C(S,t), X(S,t_J)\}$ де $t \in T_{val}$, і D_{test} – тестові дані $D_{test} = \{C(S,t), X(S,t_K)\}$де $t \in T_{test}$

Знайти модель, що прогнозує *C(S,t)*, для кадрів $t \in T_{test}$ таку, що дозволяє знизити перенавчання та мінімізувати помилку прогнозування за умов, коли (1) залежність між *C* і *X* змінюється в просторі, (2) $X(S,t_J)$ недоступний для прогнозування $C(S,t_i)$, $t_i \in T_{test}$

3.2 Базові припущення

Одним із підходів до зменшення перенавчання є пристосування наявних нейронних мереж до одного набору даних та усереднення прогнозів для кожної моделі. Подібну апроксимацію можна зробити за допомогою невеликої колекції (ансамбля) різних моделей. При необмежених обчисленнях найкращим способом «регуляції» моделі фіксованого розміру є усереднення прогнозів усіх можливих налаштувань параметрів, зважуючи кожне налаштування по його апостеріорній ймовірності з урахуванням навчальних даних [9]. В результаті, одна модель може бути використана для імітації великої кількості різних мережевих архітектур шляхом випадкового випадання вузлів під час навчання. Така процедура називається відсівом і пропонується як обчислювально дуже дешевий і надзвичайно ефективний метод регуляризації для зменшення перенавчання та зменшення помилки узагальнення в глибоких нейронних мережах усіх видів.

В основі моделі використовувалась ConvLSTM від Keras. Запропонований підхід передбачає використання моделі генерації послідовностей з

використанням ConvLSTM з регуляризацією відсіву для зменшення перенавчання та зменшення помилки узагальнення при прогнозуванні знакової мови жестів. В якості вдосконалення моделі, використано контурування і регуляризація відсіву dropout [9] для зниження перенавчання та мінімізації помилки передбачення при прогнозуванні жестів. У наступних підрозділах це розглядається більш детально.

3.3 Специфікація нейронної мережі

В цьому підрозділі дано визначення виділених ознак для розпізнавання і прогнозування жестів в відеокадрах. Кожен відеокадр має певну роздільну здатність. Довжина і ширина кадру визначається кількістю пікселів кадру, з яких формується сітка пікселів зображення одного кадру. Таким чином, зміна значення кожного пікселя відбивається на загальній сітці пікселів зображення одного кадру в момент часу *t*, і, отже, на зміну значення пікселів зображень в послідовності кадрів з плином часу. Крім того, значення кожного пікселя пов'язано з сусідніми пікселями.

Основою структури мережі для прогнозування жестів є Convolutional LSTM (ConvLSTM [10]) з 8 шарами. Її завдання полягає в обробці кожного зображення, зробленого з камери, зі швидкістю 10 кадрів в секунду для отримання значущих функцій. Крім того, вона являє собою одну з небагатьох структур, здатних масштабуватися відповідно до даних, тобто її глибину можна легко збільшити або зменшити залежно від складності сцени, без небезпеки перенавчання моделі. Розмір ядра (3, 3) підтримується на всіх шарах для поліпшення виявлення ознак у різних масштабах.

Для подальшого розширення можливостей ConvLSTM для вирішення проблеми прогнозування жестів було виконано наступні модифікації:

- на етапі попередньої обробки введено виявлення контуру, реалізованого у вигляді фільтру, який дозволяє виділити контури руки. Фільтрація на вхідному рівні дозволяє фіксувати часові залежності та зменшувати шум вхідного сигналу;

- проведена заміна 3 шарів Batch Normalization на 3 шари Dropout regularization.

На етапі прогнозування для генерації 50 відеопотоків даних та балансування просторового виводу, останній був встановлений на 6,4 Гц, відповідно до часового вікна T і частоти отримання відео 10 Гц. Усі тензори, отримані виборкою просторово-тимчасової мережі в реальному часі, розміщуються в буфері ковзного вікна, що містять до 50 10-секундних послідовностей для обробки в загальній мережі.

Просторово-часова мережа надає наступні переваги:

- стабілізує вхідні дані, виділяючи найбільш суттєві деталі для вилучення ознак;

- дозволяє отримати вищу репрезентативність отриманих ознак, запобігаючи тим самим перенавчанню моделі на етапі навчання.

Спостереження в динамічній системі в просторовій області, представленої сіткою $M \times N$, яка складається з M рядків і N стовпців, представлене наступним чином. Усередині кожної клітинки сітки є P вимірювань, які змінюються з часом. Таким чином, спостереження в будь-який час може бути представлено тензором $X \in \mathbb{R}^{P \times M \times N}$, *де* $\mathbb{R}$ позначає простір спостережуваних ознак. Якщо спостереження періодично записуються, виходить послідовність тензорів $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_t$. Завдання прогнозування просторово-часової послідовності полягає в тому, щоб передбачити найбільш ймовірну послідовність довжиною K в майбутньому, враховуючи попередні J спостережень, які включають поточні:

$$\tilde{X}_{t+1}, \ldots, \tilde{X}_{t+K} = \underset{X_{t+1}, \ldots, X_{t+K}}{argmax}\, p(X_{t+1}, \ldots, X_{t+K} | \hat{X}_{t-J+1}, \hat{X}_{t-J+2}, \ldots, \hat{X}_t)$$

Мета полягає в тому, щоб отримати $\tilde{X}_{t+1}, \ldots, \tilde{X}_{t+K}$ використовуючи мінімальну і більш ефективну топологію нейронної мережі, здатну поліпшити стандартні структури для прогнозування тимчасових дій.

3.4 Архітектура мережі для прогнозування відеопослідовностей жестів

На рис. 3.1 показана архітектура запропонованої просторово-часової ConvLSTM. Модель формується шляхом багаторазового накладення однієї нейронної мережі поверх іншої. Це означає, що вихідні дані одного шару є вхідними даними для наступного шару.

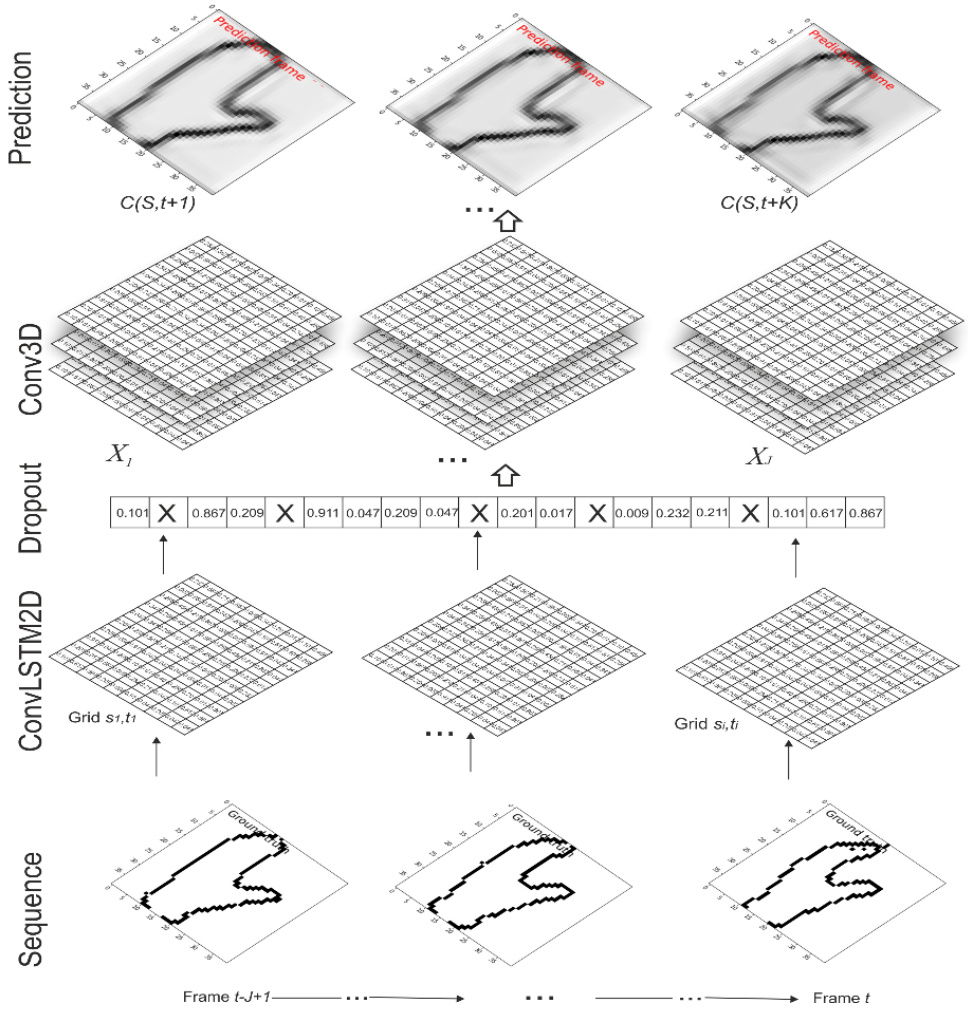


Рисунок 3.1 - Архітектура запропонованої мережі

Мережа ConvLSTM, розширює повністю підключений рівень LSTM, щоб використовувати згорткові структури як при переході між входами і станами, так і між станами, що надає можливість одночасно вивчати просторові і тимчасові відносини [10]. Структура мережі ConvLSTM, запропонована в цьому

дослідженні (рис. 3.2), заснована на реалізованій ConvLSTM в Keras з Tensorflow [2]. Основною перевагою відео в порівнянні з даними зображення є той факт, що воно пропонує набагато ширший простір для розподілу даних, додаючи часовий вимір поряд з просторовим. Цей факт обумовлює актуальність використання ConvLSTM, яка враховує саме просторово-часовий вимір.

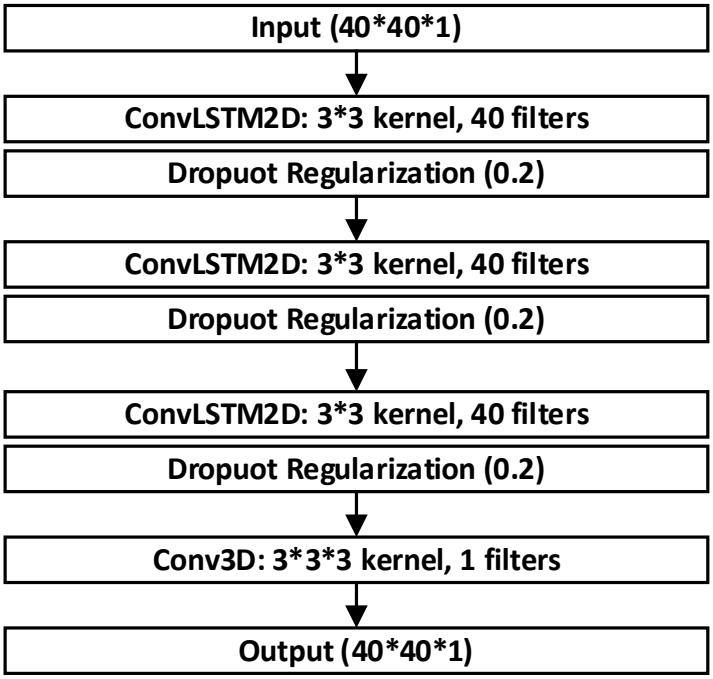


Рисунок 3.2 – Структура запропонованої ConvLSTM

Струтура запропонованої ConvLSTM (рис. 3.2) складається з чотирьох шарів ConvLSTM2D, одного шару Conv3D та трьох шарів регуляризації виключення.

Шар ConvLSTM2D майже такий самий, як і шар традиційної LSTM [5] у всіх розумінням, за винятком того, що вхідне та рекурентне перетворення є двовимірними згортковими перетвореннями. Це дає моделі можливість інтерпретувати візуальну інформацію у вигляді двовимірних зображень і здатність розуміти тимчасові послідовності, що робить її придатною для передбачення/генерування відеокадрів. ConvLSTM2D представляє спеціальну архітектуру, яка поєднує в собі функціонування LSTM із двовимірними згортками. Архітектура є повторюваною: вона зберігає прихований стан між

кроками (рис. 3.1). 2D-дані розглядаються як дані просторового виміру, які в основному захоплюються операціями згортки. 2D дані передаються в 3D мережу, яка відповідає, в основному, за часовий вимір, тимчасові характеристики цього виміру фіксуються структурою LSTM.

Шар Conv3D використовується для вилучення важливих візуальних функцій з виходів шарів ConvLSTM2D. Conv3D - це, в основному, згорковий шар, який бере карти тривимірних об'єктів і використовує тривимірні ядра для створення нових карт тривимірних об'єктів, що містять низькорівневі об'єкти, витягнуті з його вхідних даних. Шари регуляризації виключенням [3] розташовані між шарами ConvLSTM2D і шаром Conv3D. Виключення працює шляхом імовірнісного видалення або виключення входів до шару, які можуть бути вхідними змінними у вибірці даних або активаціями з попереднього шару. Це має ефект моделювання великої кількості мереж з різною мережевою структурою і, в свою чергу, робить вузли в мережі загалом більш надійними щодо входів. Останній шар Conv3D створює ядро тривимірної згортки, згорнуте з вхідним шаром по тимчасовому виміру для отримання вихідних часових рядів.

3.5 Технологія розпізнавання та прогнозування жестів

Запропонований підхід представлений чотирма етапами:

Eman 1: Контурування (фільтрація).

На цьому етапі виконується обробка вхідного зображення для виділення вузлів зображення.

Eman 2: Виключення.

Операція виключення виконується з метою зменшення перенавчання моделі за рахунок використання відсіву.

Eman 3: Класифікація

На етапі класифікації будується розподіл ймовірностей $P(\hat{y})$ сигналів, що виробляються функцією на трьох шарах ConvLSTM2D та одному шарі Conv3D.

Етап 4: Оцінювання моделі.

Останнім кроком є оцінювання результати прогнозування.

3.5.1 Фільтрація

Процес побудови контурів можна розбити на наступні етапи [1].

Застосування фільтру Гауса для згладжування зображення з метою видалення шуму. Для згладжування зображення ядро фільтра Гаусса згортається із зображенням. Цей крок трохи згладить зображення, щоб зменшити вплив очевидного шуму на детектор контурів. Рівняння для ядра фільтру Гауса розміром (2k + 1) × (2k + 1) задається наступним чином (3.1).

$$H_{ij} = \frac{1}{2\pi\sigma^2}\exp\left(\frac{i-(k+1)^2+(j-(k+1)^2}{2\sigma^2}\right);1 \le i,\ j \le (2k+1)\,. \qquad (3.1)$$

Знаходження градієнтів інтенсивності зображення.

Використовується оператор Собеля [8], що обчислює значення градієнта яскравості. Оператор Собеля використовує два квадратних ядра згортки, які оцінюють градієнт в горизонтальному і вертикальному напрямках. Після проходження згортки напрямок градієнта обчислюється як (3.2).

$$\Theta = \arctan\left(\frac{G_y}{G_x}\right), \qquad (3.2)$$

де G_y і G_x – значення для першої похідної відповідно в горизонтальному і вертикальному положеннях.

Застосування не максимального придушення, щоб позбутися помилкової реакції на виявлення контуру.

Краями визнаються пікселі, в яких досягається локальний максимум градієнта в напрямку вектору градієнта. Значення кожного пікселя невизнаного максимумом встановлюється в нуль. В результаті виходить тонка лінія контуру.

Застосування подвійного порогу для визначення потенційних контурів.

Для оцінки того, чи дійсно має місце край в конкретній точці зображення, використовуються два порога. Якщо значення пікселя відноситься вище порога, то край визнається достовірним. Інакше відкидається. Проміжним пікселям присвоюється середнє значення.

Відстеження краю за допомогою гістерезису. Виявлення контурів завершується придушенням всіх інших ребер, які є слабкими і не зв'язані з сильними краями. Пікселі вище визначеного порогового значення є крайовими пікселями. Пікселі, які межують з крайовим пікселем, і при цьому мають значення вище, ніж друге визначене значення також відносяться до групи крайових пікселів.

3.5.2 Регуляризація виключенням

Описана модель виключення [9] використана для ConvLSTM. Розглянемо нейронну мережу з L прихованими шарами. Нехай $l \in \{1,\ldots, L\}$ позначає приховані шари мережі. Нехай $z^{(l)}$ позначає вектор входів в шар l, $y^{(l)}$ позначає вектор виходів з шару l ($y^{(0)} = x$ є входом). $W^{(l)}$ і $b^{(l)}$ - ваги і зміщення в шарі l. Операція прямого зв'язку стандартної нейронної мережі може бути описана як (для $l \in \{0,\ldots, L-1\}$ і будь-якої прихованої одиниці i).

$$z_i^{(l+1)} = w_i^{(l+1)} y^l + b_i^{(l+1)},$$
$$y_i^{(l+1)} = f(z_i^{(l+1)}),$$

де f - будь-яка функція активації, наприклад, $f(x) = 1/(1+\exp(-x))$.

З виключенням, операція прямого зв'язку стає $r_j^{(l)} \sim Bernoulli(p)$

$$\tilde{y}^{(l)} = r^{(l)} \circ y^{(l)}$$
$$z_i^{(l+1)} = w_i^{(l+1)} \tilde{y}^l + b_i^{(l+1)}$$
$$y_i^{(l+1)} = f(z_i^{(l+1)})$$

Де ○ позначає поелементний добуток, також відомий як добуток Адамара. Для будь-якого шару *l*, $r^{(l)}$ є вектором незалежних випадкових величин Бернуллі, кожна з яких має ймовірність *p*, рівну 1. Цей вектор дискретизуется і множиться поелементно на виходи цього шару $y^{(l)}$, щоб створити зріджені виходи $\tilde{y}^{(l)}$. Зріджені виходи потім використовуються в якості вхідних даних для наступного шару. Цей процес застосовується на кожному шарі. Це становить вибірку підмережі з більшої мережі. Для навчання похідні функції втрат назад поширюються через мережу. Під час тесту ваги масштабуються як $W_{test}^{l} = pW^{(i)}$. Отримана нейронна мережа використовується без відсіву.

3.5.3 Прогнозування

Для згорткової LSTM [10] позначимо всі входи $X_1, \ldots, X_t$ виходи осередку $C_1, \ldots, C_t$, приховані стани $L_1, \ldots, L_t$ та вихід i_t, f_t, o_t ConvLSTM - це тривимірні тензори, останні два виміри яких є просторовими вимірами (рядки і стовпці). Щоб отримати кращу картину входів і станів, ми можемо представити їх як вектори, які стоять на просторовій сітці. ConvLSTM визначає майбутній стан певної комірки в сітці по входах і минулим станів її локальних сусідів.

Для цього використовується оператор згортки в переходах між станами і входом в стан. Ключові рівняння ConvLSTM показані в (3) нижче, де * позначаєє оператор згортки, а ○ – добуток Адамара.

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * L_{t-1} + W_{ci} \circ C_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf} * X_t + W_{hf} * L_{t-1} + W_{cf} \circ C_{t-1} + b_f)$$
$$C_t = f_t \circ C_{t-1} + i_t \circ tanh(W_{xC} * X_t + W_{hC} * L_{t-1} + b_C)$$
$$o_t = \sigma(W_{xo} * X_t + W_{ho} * L_{t-1} + W_{co} \circ C_{t-1} + b_o)$$
$$L_t = o_t \circ tanh(C_t)$$

де W - ваги, а b - зміщення.

Для вирішення завдання прогнозування просторово-часової послідовності використовується структура, яка складається з двох мереж, мережі кодування і мережі прогнозування. Початкові стани і вихідні дані комірок мережі прогнозування копіюються з останнього стану мережі кодування. Обидві мережі формуються шляхом об'єднання кількох шарів ConvLSTM. Оскільки мета прогнозування має ту ж розмірність, що і вхідні дані, об'єднуються всі стани в мережі прогнозування і подаються в згортковий шар 1 × 1 для генерації остаточного прогнозу.

Кодуюча LSTM стискає всю вхідну послідовність в тензор прихованого стану, і LSTM прогнозування розкриває цей прихований стан, щоб дати остаточний прогноз.

$$\tilde{X}_{t+1},\ldots,\tilde{X}_{t+K} = \underset{X_{t+1},\ldots,X_{t+K}}{argmax}\, p(X_{t+1},\ldots,X_{t+K}|\hat{X}_{t-J+1},\hat{X}_{t-J+2},\ldots,\hat{X}_t)$$

$$\approx \underset{X_{t+1},\ldots,X_{t+K}}{argmax}\, p(X_{t+1},\ldots,X_{t+K}|f_{encoding}(\hat{X}_{t-J+1},\hat{X}_{t-J+2},\ldots,\hat{X}_t)$$

$$\approx g_{encoding}(f_{encoding}(\hat{X}_{t-J+1},\hat{X}_{t-J+2},\ldots,\hat{X}_t))$$

Всі вхідні і вихідні елементи є тривимірними тензорами, які зберігають всю просторову інформацію.

Оцінка моделі прогнозування проводиться з використанням перехресної бінарної ентропії, точності, евклидова, манхеттенського відстаней, нормалізованої перехресної кореляції. [7]

Функція втрат обчислює перехресну бінарну ентропію між прогнозованим $\hat{X}$ та очікуваним X значенням (3.3).

$$H(X,\hat{X}) = -\sum_{i=0}^{k}\sum_{j=0}^{C}(X_{ij}\, log(\hat{X}_{ij})), \tag{3.3}$$

з k та C що є тимчасовою довжиною для поточної партії та кількості класів відповідно.

Точність [4] є метрикою, що створює дві локальні змінні, *total* і *count*, які використовуються для обчислення частоти, з якою $\hat{X}$ відповідає X. Ця частота в кінцевому підсумку повертається як бінарна точність: ідемпотентна операція, яка просто ділить загальну *total* на *count*.

Евклидова відстань розраховується між точками $X, \hat{X}$ наступним чином (3.4).

$$d_E(X, \hat{X}) = \sqrt{\sum_{i=1}^{n}(X_i - \widehat{X_i})^2}. \quad (3.4)$$

Манхеттенська відстань являє собою модуль суми різниць між реальними і прогнозованими точками даних, як це представлено нижче (3.5).

$$d_M(X, \hat{X}) = \sum_{i=1}^{n} |X_i - \widehat{X_i}|. \quad (3.5)$$

Нормалізована перехресна кореляція (NCC) [6] вимірює схожість двох кадрів зображення як функцію зміщення одного щодо іншого. Це може бути математично визначено як (3.6).

$$NCC(f, g) = \sum_{X,\hat{X}} \frac{(f(X)-\mu_f)(g(\hat{X})-\mu_g)}{\sigma_f \sigma_g}, \quad (3.6)$$

де $f(X)$ під-зображення, $g(\hat{X})$ є шаблоном, який потрібно зіставити, μ_f, μ_g позначає середнє значення під-зображення та шаблону відповідно та σ_f, σ_g позначає стандартне відхилення f та g відповідно.

Результати тестування мережі та порівняння з базовою ConvLSTM наведено у Розділі 5.

Висновки до третього розділу

1. Дана глава детально розглядає пропоновану модель прогнозування жестів рук у відеопослідовностях на основі згорткової LSTM.

2. Наведено архітектуру розробленої просторово-часової нейронної мережи. Запропоновано використовувати у мережі даного типу регуляризацію виключенням dropout та контурування зображення за методом Canny.

3. Дістала подальшого розвитку технологія розпізнавання та прогнозування жестів на основі моделі генерації послідовностей з використанням ConvLSTM.

4. Для подальшого розширення можливостей ConvLSTM для вирішення проблеми прогнозування жестів було виконано наступні модифікації: (1) на етапі попередньої обробки введено виявлення контуру, реалізованого у вигляді фільтру, що дозволяє виділити контури руки. Фільтрація на вхідному рівні дозволяє фіксувати часові залежності та зменшувати шум вхідного сигналу; (2) проведена заміна 3 шарів batch normalization на 3 слоя dropout regularization що дозволило знизити перенавчання та мінімізувати помилки передбачення при прогнозуванні жестів.

5. Отримана точність склала 90%, що на 30% краще ніж у моделі ConvLSTM з batch normalization (60%).

Результати розділу опубліковано в роботах автора [1, 8] (Додаток А)

Література до третього розділу

1. Canny J. (1986) A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, vol. 6, pp. 679-698.

2. Chollet F., et al., (2015) Keras. Режим доступу: https://github.com/fchollet/keras (12.11.2020).

3. Hinton G.E., Srivastava N., Krizhevsky A., Sutskever I., Salakhutdinov R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *ArXiv, abs/1207.0580*.

4. Keras: Accuracy metrics. Режим доступу: https://keras.io/api/metrics/accuracy_metrics/#accuracy-class (21.09.2020).

5. LeCun Y., Bengio Y. (1995) Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks,* 3361(10), p.1995

6. Prateep B., Das S. (2017) Temporal Coherency based Criteria for Predicting Video Frames using Deep Multi-stage Generative Adversarial Networks. *NIPS* (2017).

7. Godoy D. Understanding binary cross-entropy / log loss: a visual explanation. Режим доступу: https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a (21.09.2020).

8. Sobel I., Feldman G. (1968) A 3x3 isotropic gradient operator for image processing. *A talk at the Stanford Artificial Project*, pp. 271-272.

9. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* Vol. 15, 1, pp. 1929-1958.

10. Xingjian S.H.I., et al. (2015) Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, pp. 802-810.

11. Yuan Z., Zhou X., Yang T. (2018) Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). Association for Computing Machinery*, New York, NY, USA, pp. 984-992. DOI:https://doi.org/10.1145/3219819.3219922.

РОЗДІЛ 4
РОЗРОБЛЕННЯ ТЕХНОЛОГІЇ ВЗАЄМОДІЇ МІЖ ХІРУРГОМ ТА СИСТЕМОЮ ПЕРЕГЛЯДУ ЗОБРАЖЕНЬ В ОПЕРАЦІЙНІЙ ЗАЛІ НА ОСНОВІ МОДЕЛІ СКІНЧЕННОГО АВТОМАТУ

У розділі представлено новий набір даних динамічних жестів і визначено лексикон для навігації медичних зображень в операційній залі. Розроблено технологію взаємодії між хірургом та системою перегляду зображень в операційній залі на основі скінченного автомату (FSM). Надано опис реалізації ЛМВ за допомогою жестів на основі FSM з пороговою моделлю, яка використовує порогову ймовірність за введеним шаблоном і обчислює подібність між двома ключовими кадрами за алгоритмом відстані Хаусдорфа для управління переглядом зображень в операційній залі.

4.1 Датасет для безконтактного перегляду медичних зображень

Метою нового датасету було створити просту, але виразну структуру, яку хірурги можуть легко інтерпретувати та прийняти для взаємодії зі спеціалізованим безконтактним хірургічним монітором в ОЗ без запам'ятовування складних мовних правил. В найкращому випадку, природність інтерфейсу ЛМВ вимагає, щоб кожен жест, який виконується користувачем, був інтерпретованим [8], але сучасний рівень технологій розпізнавання жестів на основі зору поки ще не в змозі забезпечити повне вирішення цієї проблеми. З огляду на це зауваження, в дисертації основна увага приділяється використанню жестів рук, які цілеспрямовано даються у вигляді інструкцій, і обмежені навмисними, виразними рухами. Тобто вважається, що пози рук і конкретні жести використовуються як команди на мові команд.

Фактично це означає, що жести не обов’язково мають бути природними, але можуть бути розроблені відповідно до ситуації або на основі стандартної мови жестів. Так, наприклад, під час операції хірургу можуть знадобитися деталі про структурні особливості частини тіла пацієнта або детальну модель органу,

щоб скоротити час операції або підвищити точність її результату. Це досягається шляхом взаємодії з різноманітними системами візуалізації, такими як МРТ, КТ, рентгенівська система, які збирають дані пацієнта та відображають їх на екрані у вигляді детального зображення. Хірург може реалізувати взаємодію з зображеннями, виконуючи жести рук перед камерою, використовуючи технології комп'ютерного зору. За допомогою жестів можна виконувати масштабування, обертання, обрізання зображення та перехід до наступного або попереднього слайда без використання будь-яких інших периферійних пристроїв, таких як миша, клавіатура або сенсорний екран. Тому було обрано невелику колекцію візуально відмітних та інтуїтивних жестів, які б покращили ймовірність надійного розпізнавання жестів. Взаємодія хірургів зі спеціалізованим безконтактним хірургічним монітором в ОЗ передбачає розпізнавання та реагування щонайменше на десять жестів, перелічених у [4].

Реалізація сценарію безконтактного управління переглядом медичних зображень в ОЗ передбачала (1) створення спеціалізованого словникового запасу жестів, (2) відеозапис та використання набору жестів рук.

4.1.1 Створення словникового запасу жестів

З метою визначення найбільш інтуїтивних та легких у виконання жестів, спочатку, хірургам був запропонований базовий набір [7], за допомогою якого пропонувалося дати опис найпоширеніших дій, які вони виконують з медичними зображеннями під час операцій для навігації зображень, зміни яскравості, орієнтації та управління масштабом тощо.

Всі запропоновані жести оцінювалися за чотирибальною шкалою (0 – «категорично не прийнятне», 1 – «не прийнятне», 2 – «нейтрально», 3 – «прийнятне») за наступними критеріями [2]:

- інтуїтивність (жест інтуїтивно зрозумілий і відповідає виконуваній команді),
- простота використання (жест легко і зручно виконувати),

- легкість запам’ятовування.

В результаті, всі жести рук з набору [7] отримали середню оцінку між 2 і 3 («нейтрально» і «прийнятне»). Більшість хірургів не дали пропозицій щодо покращення набору, що дало змогу зробити висновок, що базовий набір цілком прийнятний.

Кілька жестів руками, особливо маніпуляцій з яскравістю та орієнтацією зображення, отримали відносно нижчий рейтинг за всіма критеріями.

Хірурги, що приймали участь в опитуванні пропонували віддати перевагу жестам, що включають менш складні рухи. Щоб визначити остаточний набір жестів для навігація зображень в ОЗ, була використана процедура, запропонована в [9]:

1. Якщо жест отримав стабільно нижчі бали за всіма критеріями, він замінювався на: (a) запропонований покращений жест за згодою в оцінці принаймні двома хірургами; (b) запропонований покращений жест, наданий принаймні одним хірургом [5]

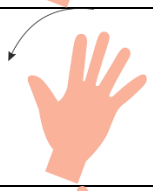
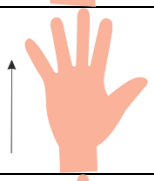
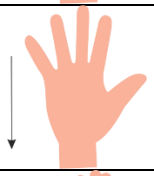
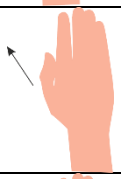
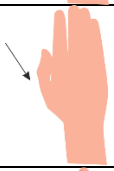
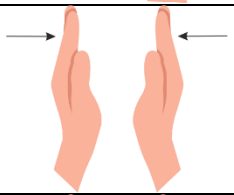
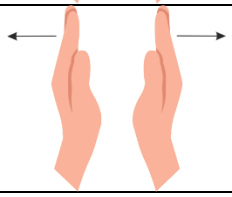
2. Якщо пропозицій немає або якщо жест отримав задовільні оцінки, жест залишається в наборі.

4.1.2 Опис набору динамічних жестів

Для даного дослідження було створено власний набір даних який містить 10 видів динамічних жестів рук для безконтактного управління переглядом медичних зображень. Ці жести руками та їх семантика показані в табл.4.1. Стрілками позначений напрям руху рук.

Жести рук 1, 2 можуть керувати напрямком повороту зображення за годинною стрілкою або проти. Жести рук 3, 4, 5, 6 можуть керувати напрямом перегляду зображення, тобто вліво, вправо, вгору та вниз. Жести рук 7, 8 можуть керувати яскравістю зображення – зменшувати або збільшувати її. Жести рук 9, 10 можуть керувати збільшенням або зменшенням розміром перегляду зображення.

Таблиця 4.1 – Жести рук для безконтактного перегляду медичних зображень (Адаптовано з [4, 7])

Номер жесту	Управління медичним зображенням	Жест
1	Поворот зображення за годинною стрілкою	
2	Поворот зображення проти годинної стрілки	
3	Перегляд вліво	
4	Перегляд вправо	
5	Перегляд вгору	
6	Перегляд вниз	
7	Збільшення яскравості зображення	
8	Зменшення яскравості зображення	
9	Зменшення розміру перегляду зображення	
10	Збільшення розміру перегляду зображення	

Розпізнаючи ці жести рук, програмний засіб дозволяє реалізувати безконтактний перегляд необхідних медичних зображень за умов збереження стерильності рук лікаря. Кадри відеопослідовностей запису жестів, описаних у табл. 4.1, надано на рис. 4.1. Ключові кадри відеопослідовностей містять перший, останній кадри жесту та один з проміжних між ними кадрів.

Рисунок 4.1 – Ключові кадри відеопослідовностей запису жестів

Керуючись консультацією хірургів, для зручності було додано також допоміжну функцію блокування / розблокування системи. Дії блокування та розблокування сигналізують системі про намір хірурга використовувати систему. Блокування системи зменшує можливість розпізнавання ненавмисних

жестів рук, що виконуються користувачем. Кількісна інформація про новий набір даних динамічних жестів надана у табл. 4.2.

Таблиця 4.2 – Характеристики набору даних динамічних жестів для реалізації ЛМВ в операційній залі (Набір 2)

Назва жесту	Кількість відеозаписів жесту	Середній час одного відеозапису (сек)	Кількість кадрів
Поворот за годинною стрілкою	12	0.715	227
Поворот проти годинної стрілки	17	0.72	324
Перегляд вліво	14	0.69	254
Перегляд вправо	11	0.815	241
Перегляд вгору	18	0.72	343
Перегляд вниз	13	1.165	421
Збільшення яскравості	18	0.695	329
Зменшення яскравості	13	0.295	82
Зменшення розміру	14	0.74	275
Збільшення розміру	15	0.785	315

4.2 Технологія взаємодії між хірургом та системою перегляду зображень в операційній залі на основі скінченного автомату

Для безпомилкової роботи системи безконтактного перегляду медичних зображень, заснованої на жестах, також слід враховувати стабільність системи. У разі реалізації неправильного жесту рук або в процесі переходу між різними жестами, або коли система тільки запускається і жест руки раптово змінюється, система безконтактного перегляду медичних зображень, заснована на жестах, може працювати нестабільно. Необхідно реалізувати виключення нестабільної інформації про жести рук і використання тільки стабільних вихідних даних жестів рук для управління переглядом медичних зображень в ОЗ.

Для вирішення цієї задачі розроблено модель скінченного автомату (finite state machine, FSM). Скінченний автомат використовується для подання кінцевих станів, переходів і дій між цими станами. Система безконтактного перегляду медичних зображень, заснована на жестах, враховується як мовна форма, яка

висловлює семантичні команди за допомогою жестів рук. Скінченний автомат є зручним для моделювання семантики різних жестів рук. Для кожного визначеного жесту руки створено модель скінченного автомату. В процесі взаємодії між хірургом і системою безконтактного перегляду медичних зображень, спочатку відеокамера збирає дані кожного кадру, що містять жест рукою. Далі використовується розпізнавання типу жесту руки та прогнозування жесту на підставі початкових кадрів відеопослідовності, а потім оброблена інформація жесту руки вводиться в відповідну модель скінченного автомата. Це дозволяє виконувати зумовлену маніпуляцію з зображеннями або відповідну функцію. Крім того, детермінована модель на основі FSM використовується для перетворення жестів в інструкції.

Напрямок руху рук хірурга моделюється як послідовність кадрів зображення. Ознаки прогнозованих кадрів відеопослідовностей зображень, що отримуються з використанням моделей розпізнавання та прогнозування кадрів відеопослідовностей жестів використовуються в якості вхідних даних для моделі FSM. Взаємодія з системою безконтактного управління переглядом медичних зображень в ОЗ відбувається в режимі реального часу з урахуванням даних, що отримуються в поточний момент часу, та на підставі яких генеруються прогнозовані дані. Для дослідження даних, що отримуються в поточний момент часу використовується контекстна інформація, що зберігається в FSM. Контекстною інформацією є результати моделювання розпізнавання та прогнозування жестів, представлені інформацією про схожість прояву того чи іншого жеста. Міра схожості жесту отримується з використанням прогнозованих ознак кадру.

Контекстна інформація системи ЛМВ в умовах ОЗ у відповідності до формального визначення скінченного автомату представлена наступним чином.

Кожен стан SM має параметри $< F_s, T_s, H_s, \sum_s, J_s >$ для визначення отриманої ним інформації, де F_s - це просторова інформація прогнозованих кадрів, тобто прогнозовані ознаки для розпізнавання кадрів відеопослідовностей, T_s - тимчасові дані прогнозованих кадрів, H_s - міра схожості кадрів, $\sum_s$ -

інформація про кількість подій в даному стані, J_s - інформація про порогові значення при визначенні події для $\sum_s$. Граничні значення встановлюються для визначення події розпізнавання кадрів з використанням значення заходи схожості кадру і для визначення кількості розпізнаних кадрів і розпізнаних жестів для переходу станів.

Розроблена модель скінченного автомату складається з трьох станів: початковий, перехідний та робочій. При отриманні нових даних *Dt* при розпізнаванні жесту за умови (4.1) виконується перехід з початкового стану на перехідний стан.

$$Dt(f_k, t_k, dt_k) \vee Dt(f_{k+1,\dots,i}, t_{k+1,\dots,i}, dt_{k+1,\dots,i}). \quad (4.1)$$

При переході з початкового стану на перехідний оновлюється інформація про кількість подій, тобто $\sum_s = 0$.

Перехід з перехідного стану на робочій стан виконується при дотриманні наступної умови [6]:

$$(Dt(f_k, t_k, dt_k) \vee D(f_{k+1,\dots,i}, t_{k+1,\dots,i}, dt_{k+1,\dots,i}) \geq J_p) \wedge \sum_{k+1,\dots,i} > H_i, \quad (4.2)$$

де H_p є пороговим значенням міри подібності для підтвердження розпізнавання прогнозованого кадру, а H_i є пороговим значенням для i послідовних кадрів з розпізнаним прогнозованим жестом. При невиконанні умови виконується перехід на початковий стан.

Передача з робочого стану у систему ЛМВ в ОЗ виконується при дотриманні умови (4.3).

$$\sum_s = H_i * H_l, \quad (4.3)$$

де J_l - є пороговим значенням для l послідовних розпізнаних жестів. Якщо жест у робочому стані має l послідовних розпізнавань жесту, цей жест

використовується для управління системою ЛМВ в умовах ОЗ. При невиконанні умови виконується перехід на початковий стан.

Схему моделі скінченного автомату представлено на рис. 4.2.

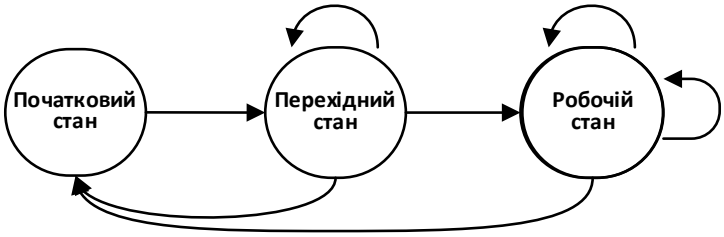


Рисунок 4.2 – Схема моделі скінченного автомату

Таким чином, запропонований FSM для розпізнавання жестів складається з кінцевої кількості ключових кадрів та відповідної тривалості ключового кадру. Тривалість кадру вимірюється через кількість відеокадрів між відповідним ключовим кадром і наступним кадром. Перехід стану відбувається лише тоді, коли виконуються подібність форми ключових кадрів та критерії тривалості.

Далі накладається додаткове обмеження, що кожен маркер жестів повинен бути виявлений протягом 10 послідовних кадрів для активації переходу. Це обмеження додає надійності, щоб запобігти пропущеній або неправильній класифікації певного жеста. Таке обмеження допомагає також відкинути помилкові жести, які можуть бути виявлені, коли хірург переходить з одного жесту на інший. Крім того, оскільки відображення є індивідуальним, малоймовірно, що буде сформовано неправильну інструкцію, навіть якщо хірург помилково виконує деякі неточні жести, оскільки в кожному стані немає інших правил переходу, крім правильних.

Для вимірювання подібності між двома ключовими кадрами використана міра відстані Хаусдорфа [3], яка визначається як максимум між двома наборами точок O і I:

$$H(O, I) = \max\{h(O, I), h(I, O)\}, \tag{4.1}$$

де $h(O, I) = \max_{o \in O} \min_{i \in I} \|o - i\|$ і $h(I, O) = \max_{i \in I} \min_{o \in O} \|i - o\|$.

Точки ознак позначаються o_1, ..., o_m для множини O та i_1, ..., i_n для множини I.

Обчислення відстані Хаусдорфа виконується за допомогою алгоритму перетворення відстані [10]. Алгоритм базується на припущенні про наявність екстремальних вершин [1], тобто що при оптимізації відстані між двома непересічними опуклими багатокутниками, точки з екстремальними відстанями завжди є вершинами. Це означає, що для знаходження відстані Хаусдорфа потрібно обчислювати лише відстань між вершинами. Крім того, встановлюється поріг J_s, який дозволяє певний ступінь дисперсії форми у кожному стані [6]. Під час навчання вхідна послідовність жестів перетворюється у послідовність станів, яка складається з ключових кадрів разом з інформацією про їх тривалість. Це дає подання конкретного жесту в узагальненому форматі.

4.2 Взаємодія розроблених компонент ЛМВ

Схему взаємодії розроблених моделей розпізнавання та прогнозування відеопослідовностей зображень жестів та моделі скінченного автомату представлено на рис. 4.3.

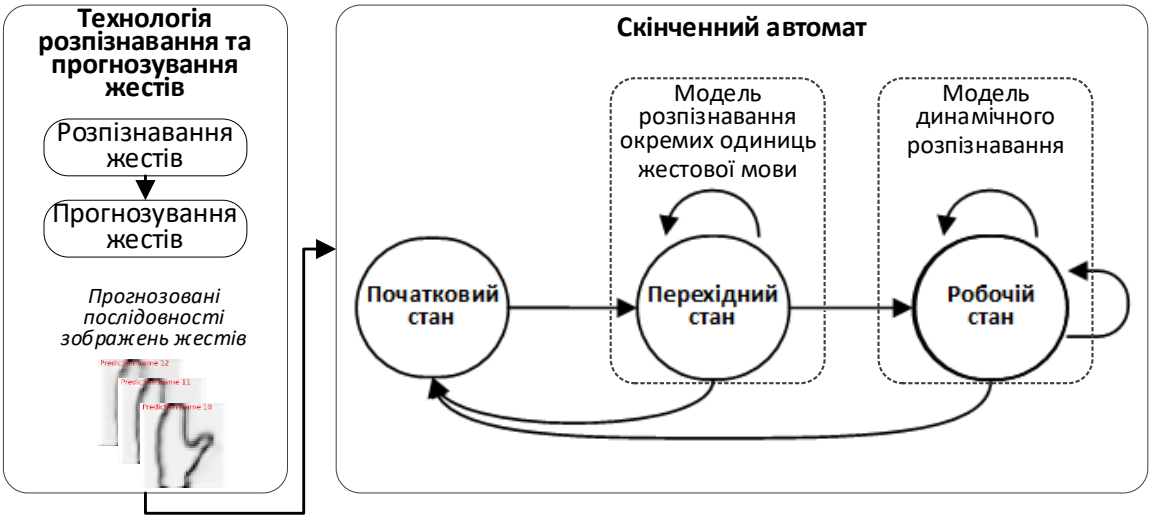


Рисунок 4.3 – Схема взаємодії розроблених компонент ЛМВ

Дані відео жестів поступають на вхід моделі прогнозування. Дані проходять всі описані вище етапи та прогнозуються послідовності зображень жестів, які є вхідними даними для скінченного автомату. Для запобігання помилкових спрацьовувань на перехідному стані моделі скінченного автомату використовується модель статичного розпізнавання жестів, а на робочому стані використовується модель динамічного розпізнавання жестів, розробка яких представлена в Розділі 2.

4.2.1 Модель скінченного автомату для безконтактного управління переглядом медичних зображень в операційній за допомогою жестів

Використання інформаційної технології для безконтактного управління переглядом медичних зображень в ОЗ відбувається з використанням розробленої моделі скінченного автомату для безконтактного управління переглядом медичних зображень.

Для кожної заздалегідь визначеної маніпуляції зображеннями, що представлено в п. 4.1, створено модель скінченного автомату. Використовувана контекстна інформація про кадри є вхідною інформацією для моделі розпізнавання та прогнозування жестів, а також, в результаті моделювання, отримується інформація про ймовірності прояву того чи іншого жестів. Міра ймовірності жесту отримується з використанням прогнозованих ознак кадру.

Процес взаємодії між хірургом та системою перегляду зображень в операційній представлено наступним чином:

1. Отримуються дані кожного кадру, що містять жести рук.

2. Метод розпізнавання та прогнозування жестів рук, що включає модель динамічного розпізнавання жестів рук, використовується для розпізнавання типу жесту руки та прогнозування жесту.

3. Оброблена інформація прогнозування кадрів жесту руки вводиться у відповідну модель скінченного автомату.

4. Системою перегляду медичних зображень виконується управління медичним зображенням.

Процес відбувається в режимі реального часу. Прогнозування кадрів жестів рук покликане зменшити час відповіді системи перегляду зображень в операційній. Необхідно враховувати коректність зіставлення жестів рук хірурга системі перегляду медичних зображень в операційній, щоб відфільтрувати інформацію про нестабільні жести рук і використовувати тільки стабільні вихідні дані для безконтактного управління переглядом зображень.

Відношення переходу між станами моделі безконтактного управління медичним зображенням за допомогою жесту 1 встановлюється з використанням моделі скінченного автомата як це показано на рис. 4.4.

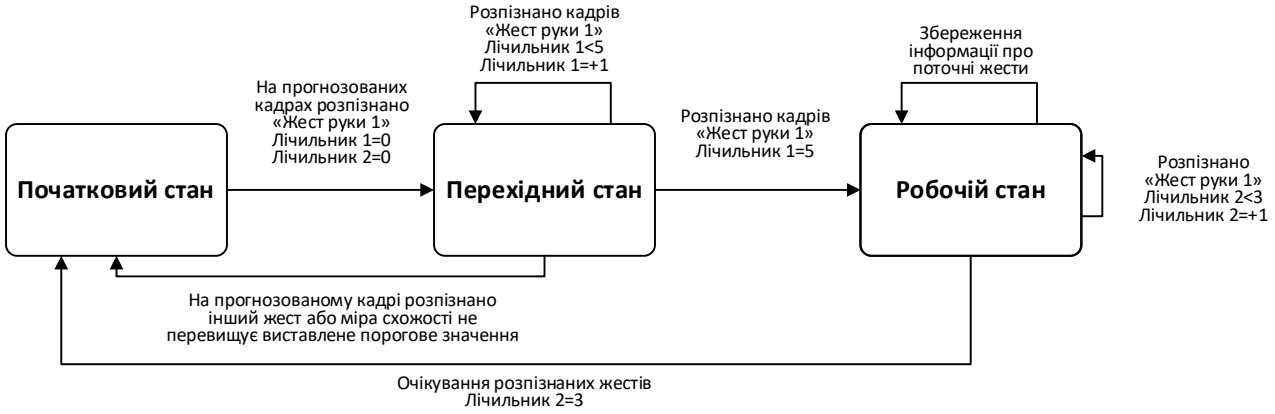


Рисунок 4.4 - Схема моделі скінченного автомату на прикладі жесту 1

Принцип реалізації моделі скінченного автомата безконтактного управління медичним зображенням за допомогою жесту 1 представлений наступним чином. Початковий час системи знаходиться в початковому стані. У цей час медичне зображення не керується. Коли хірург виконує жест рукою, модель розпізнавання та прогнозування отримує відеопослідовність кадрів з жестом, розпізнає початок жесту та прогнозує наступні кадри, система отримує прогнозовані кадри та розпізнає, що прогнозований кадр з жестом є «Жест руки 1», вона переходить в перехідний стан і скидає лічильник 1 і лічильник 2. У цей час зображення ще не управляється. Проводиться підрахунок послідовних прогнозованих кадрів, що перевищують виставлене порогове значення міри

схожості для «Жест руки 1». Якщо в цьому стані розпізнається інший жест руки, крім «Жест руки 1», або міра схожості не перевищує виставлене порогове значення, система повертається в початковий стан. Якщо інший жест не розпізнається і «Жест руки 1» у перехідному стані система має п’ять розпізнаних послідовних прогнозованих кадрів, що перевищують виставлене порогове значення міри схожості, здійснюється перехід у робочий стан і передається інформація про розпізнаний прогнозовані кадри. У робочому стані зберігається інформація про поточний жест та підраховується кількість розпізнаних жестів. Якщо «Жест руки 1» у робочому стані має три послідовних співпвдання з контекстом жесту, цей жест використовується для управління медичним зображенням.

4.2.2 Принцип використання контексту для ЛМВ

Результат розпізнавання та прогнозування, отримуваний від останнього шару нейронної мережі будь-якої моделі класифікації жестів складається з двох параметрів (d_j, α), де d_j – клас розпізнаного жесту, $d_j \in D$, $D = \{d_1, d_2, \ldots, d_n\}$, α – confidence або оцінка довіри. Оцінки довіри можуть бути використані як інтерпретовані ймовірності, які подаються на наступний етап системи прийняття рішень, або як значення, які визначають, коли не діяти згідно з прогнозами нейронної мережі у порівнянні з порогом. В якості системи прийняття рішень використана FSM модель.

Система ЛМВ в поточному контексті або стані спілкується з FSM, щоб дізнатись, який жест слід ігнорувати чи передавати на виконання. Завдяки наперед визначеному значенню α, на вхід FSM приймаються лише правильні жести, і FSM може перейти до наступного стану. Наприклад, модель FSM на рис. 4.5, має шість контекстів в режимі «перегляд вліво-вправо», представлених вузлом «перегляд стека / пересування». Для кожного вузла існує конкретна піддія або жест, обмежений FSM. Таким чином, при відтворенні стану відео, враховуються лише жести, якими можна користуватися. На рис. 4.5 показано

приклад того, як контекстно-орієнтована модель працює при виборі жестів в режимах «перегляд стека / пересування» і «зміна масштабу».

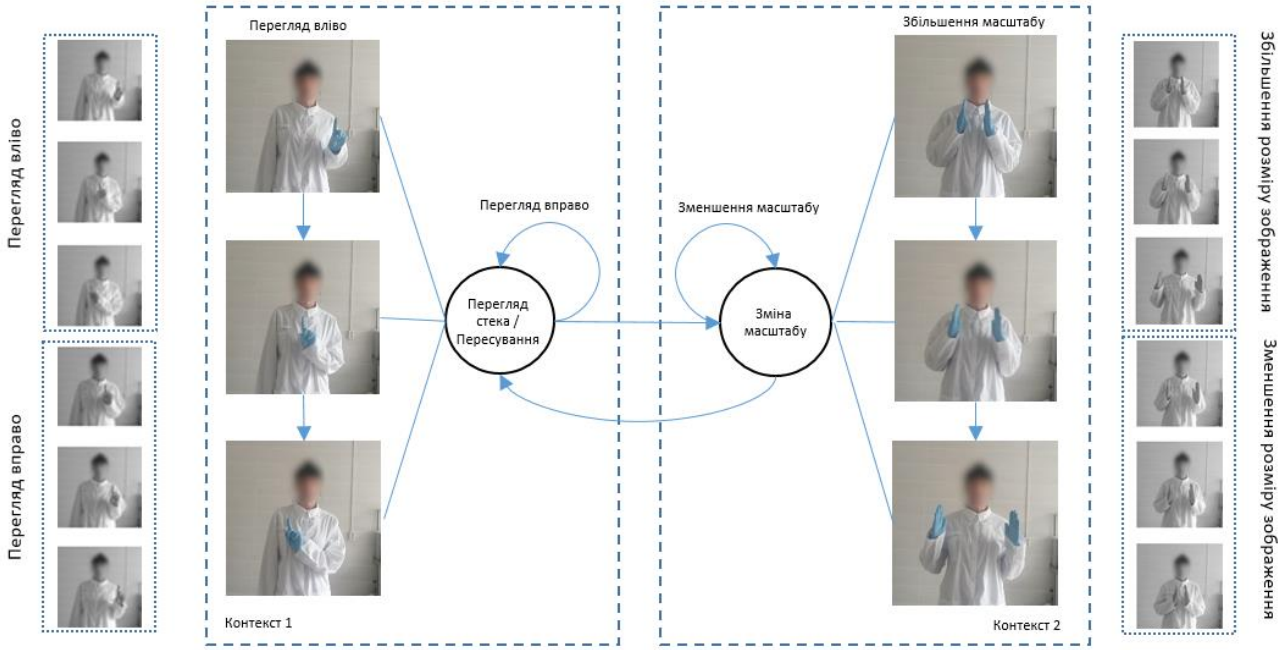


Рисунок 4.5 - Приклад реалізації двох контекстів у FSM моделі перегляду двовимірного зображення

На відміну від інших алгоритмів узгодження шаблонів, де послідовність зображень спочатку перетворюється на статичний шаблон, а потім порівнює її з попередньо збереженими прототипами дій під час розпізнавання, програма розпізнавання FSM порівнює лише вибрані ключові кадри вхідної відеопослідовності з існуючими ключовими кадрами у FSM. Порівняння форм на основі ключового кадру значно підвищує швидкість розпізнавання. Результати тестування моделі надано у п. 5.3.4.

Висновки до четвертого розділу

1. Представлено новий набір даних динамічних жестів і визначено лексикон для навігації медичних зображень в операційній залі. Метою нового датасету було створити просту, виразну структуру, яку хірурги можуть легко інтерпретувати та прийняти для взаємодії зі спеціалізованим безконтактним хірургічним монітором в ОЗ без запам'ятовування складних мовних правил.

2. Набір даних містить 10 видів динамічних жестів рук і дозволяє реалізувати безконтактний перегляд необхідних медичних зображень за умов збереження стерильності рук лікаря.

3. Розроблено технологію взаємодії між хірургом та системою перегляду зображень в операційній залі на основі скінченного автомату (FSM).

4. Удосконалено модель скінченного автомату для безконтактного управління переглядом медичних зображень за допомогою жестів яка, на відміну від існуючих, використовує дані прогнозованих кадрів відеопослідовностей, що дозволяє зменшити час відгуку системи.

5. Взаємодія з системою безконтактного управління переглядом медичних зображень в ОЗ відбувається в режимі реального часу з урахуванням даних, що отримуються в поточний момент часу, та на підставі яких генеруються прогнозовані дані.

Результати розділу опубліковано в роботах автора [7, 16] (Додаток А)

Література до четвертого розділу

1. Borgefors G. (1986). Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, vol. 34(3), pp. 344-371. doi:10.1016/s0734-189x(86)80047-0

2. Hurstel A., Bechmann D. (2019) Approach for Intuitive and Touchless Interaction in the Operating Room. *J* 2: 50– 64. doi:10.3390/j2010005.

3. Huttenlocher D.P., Klanderman G.A., Rucklidge W.J. (1993) Comparing Images using the Hausdorff Distance, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 15 (9), pp. 850-863.

4. Jacob M.G., Wachs J.P., Packer R.A. (2013) Hand-gesture-based sterile interface for the operating room using contextual cues for the navigation of radiological images. *J Am Med Inform Assoc* 20(e1): e183-186.

5. Jacob M. G., Wachs J. P. (2014). Context-based hand gesture recognition for the operating room. *Pattern Recognition Letters*, vol. 36, pp. 196-203. doi:10.1016/j.patrec.2013.05.024

6. Lee H.-K., Kim J.H. (1999). An HMM-based threshold model approach for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(10), 961–973.* doi:10.1109/34.799904

7. Madapana N., Gonzalez G., Rodgers R., Zhang L., Wachs J. P. (2018). Gestures for Picture Archiving and Communication Systems (PACS) operation in the operating room: Is there any standard?. *PloS one*, 13(6), e0198092. https://doi.org/10.1371/journal.pone.0198092.

8. Pavlovic V.I., Sharma R., Huang, T.S. (1997) Visual Interpretation of Hand Gestures for HumanComputer Interaction: A Review. In *IEEE Transactions on Pattern Analysis and Machine Intelligence.* Vol. 19 (7) pp. 677-695.

9. Salvador R.A., Naval Jr.P. Towards a Feasible Hand Gesture Recognition System as Sterile Interface in the Operating Room with 3D Convolutional Neural Network. Режим доступу: https://www.academia.edu/44021842/Towards_a_Feasible_Hand_Gesture_Recognition_System_as_Sterile_Interface_in_the_Operating_Room_with_3D_Convolutional_Neural_Network (23.11.2020).

10. Taha A.A., Hanbury A. (2015) An efficient algorithm for calculating the exact Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37(11), pp. 2153-2163. DOI: 10.1109/tpami.2015.2408351.

РОЗДІЛ 5
РОЗРОБЛЕННЯ ТА ПРАКТИЧНЕ ВИКОРИСТАННЯ ЕЛЕМЕНТІВ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ЛЮДИНО-МАШИННОЇ ВЗАЄМОДІЇ З ВИКОРИСТАННЯМ ЖЕСТІВ

У розділі представлено елементи інформаційної технології людино-машинної взаємодії в умовах операційної зали з використанням жестів. Розкрито етапи розробки та реалізації інформаційної технології у вигляді системи безконтактного перегляду зображень, представлено процес взаємодії з системою в режимі реального часу, реалізацію моделі скінченного автомату, наведено результати реалізації та оцінки адекватності моделей та елементів інформаційної технології.

5.1 Елементи інформаційної технології безконтактного управління медичними зображеннями в операційній залі з використанням жестів

Як зазначалося в попередніх розділах, для реалізації ЛМВ і безконтактного управління медичними зображеннями в операційній залі з використанням жестів необхідна реалізація динамічного процесу розпізнавання жестів, для якого досить складно вивчити як просторові, так і часові особливості за допомогою лише одного методу вилучення ознак [11]. Для вирішення цієї проблеми в дисертаційній роботі пропонується три різні моделі і технологія, що складається з декількох процесів, таких як збір даних, попередня обробка даних, навчання та тестування моделей.

У цьому підрозділі пояснюються основні етапи розробки запропонованої інформаційної технології, яка забезпечує процеси розробки моделей розпізнавання жестів рук, що базуються на комп’ютерному зорі та створює концептуальну основу для розробки системи людино-машинної взаємодії з використанням жестів в режимі реального часу.

5.1.1 Базові елементи системи безконтактного перегляду медичних зображень, заснованої на жестах

Розроблені моделі та метод є складовими компонентами інформаційної технології ЛМВ в умовах ОЗ з використанням жестів. За високого рівня абстракції розроблена інформаційна технологія має наступну структуру, представлену на рис. 5.1.

Рисунок 5.1 – Базові елементи системи людино-машинної взаємодії, заснованої на жестах

Для отримання та збору кадрів відеопослідовностей жестів рук використовується відеокамера, яка фіксує рухи жестів рук у тривимірному просторі в режимі реального часу та передає їх на модель розпізнавання та прогнозування жестів рук, побудованої на основі глибокої нейронної мережі для прогнозування відповідних жестів рук на підставі початкового руху руки. Після цього прогнозовані кадри передаються в систему управління переглядом медичних зображень, тим самим дозволяючи хірургу контролювати та управляти переглядом медичних зображень.

Розроблена система ЛМВ на основі жестів працює з фіксованими кадрами в секунду. Кожного разу, коли зображення знімається камерою, система працює наступним чином:

1) Зняте зображення попередньо обробляється, і детектор рук намагається відфільтрувати зображення рук із захопленого зображення; весь процес завершується, якщо нічого не виявлено.

2) Класифікатор CNN використовується для розпізнавання жестів з обробленого зображення.

3) Результати розпізнавання та оцінки подаються до центру контролю; використовується проста імовірнісна модель, щоб вирішити, яку відповідь повинна дати система.

Для реалізації інформаційної технології ЛМВ в умовах ОЗ розглянуто сценарій безконтактного управління переглядом медичних зображень, що містить використання розроблених моделей розпізнавання окремих одиниць жестової мови, моделі динамічного розпізнавання жестів та методу розпізнавання та прогнозування жестів. Розроблена інформаційна технологія також передбачає розробку та використання моделі скінченного автомату в умовах операційної для вирішення завдань безконтактного управління переглядом медичних зображень.

5.1.2 Структура інформаційної технології забезпечення ЛМВ з використанням жестів

Елементи інформаційної технології представлені у табл. 5.1 згідно з алгоритмом її реалізації. Структура інформаційної технології ЛМВ з використанням жестів складається з наступних процесів: визначення мети ЛМВ (управління переглядом зображень, взаємодія з роботом, віртуальна реальність), інтерфейсів управління жестами (веб-камера, Microsoft Kinect, рукавички); об’єктів взаємодії (робот, монітор), типів взаємодії (одностороння - монітори та дисплеї, двостороння - робот), словникового запасу жестів; формування вимог ТЗ до ЛМВ на основі жестів; визначення базових компонент ЛМВ на основі жестів, важливих для конкретного застосування (розпізнавання статичних жестів, розпізнавання динамічних жестів, прогнозування жестів); визначення моделей і нейромережевих архітектур, використовуваних для розпізнавання жестів; Визначення правил взаємодії ЛМВ на основі жестів; визначення стратегії спільного використання моделей розпізнавання жестів, та їх налаштування для відповідних інтерфейсів управління жестами та потенційних застосувань в системах ЛМВ.

Таблиця 5.1 – Елементи інформаційної технології

№	Процес	Вхідні дані	Вихідні дані	Ресурси	Керування
1	Збір інформації та передпроектний аналіз (мета, інтерфейс управління; об'єкти взаємодії, тип взаємодії)	- вимоги замовника; - характеристик и засобів взаємодії, введення та виведення інформації	- передпроектний перелік цілей, вимог та умов використання; - технічний звіт; - проектне завдання	-ЕПР; -Керівники проекту; -БД нормативних -документів, -стандартів, -інструкцій	система чинників, що впливають на проектне рішення.
2	Формування вимог ТЗ до ЛМВ на основі жестів	-технічне завдання; -умови використання; -словник жестів	- вимоги до системи; - розподіл задач; - показники успішності	- нормативні документи; - стандарти; - вимоги до ЛМВ на основі жестів	нормативні документи, стандарти, інструкції
3	Визначення базових компонент ЛМВ на основі жестів важливих для конкретного застосування	дизайн-концепт розпізнавання статичних жестів, розпізнавання динамічних жестів, прогнозування жестів	- правила та прийоми; - словник жестів; - набір правил відображення; - набір інструкцій	БД компонент ЛМВ на основі жестів	типологія базових компонентів
4	Визначення моделей і нейромережевих архітектур, використовуван их для розпізнавання жестів	-архітектурна ситуація	поточна верифікація	БД моделей	досвід проектування аналогічних систем: аналоги, прототипи, типові архітектури; -налаштування параметрів мережі та обробки даних
5	Визначення правил взаємодії ЛМВ на основі жестів	- словник жестів; - набір правил відображення; - набір інструкцій	згенерована інструкція ЛМВ	- набір інструкцій; - інформаційне забезпечення для оцінювання розпізнавання жесту	комплекс моделей розпізнавання жестів, технологія розпізнавання та прогнозування жестів на основі моделі генерації послідовностей
6	Визначення стратегії спільного використання моделей, їх налаштування для відповідних інтерфейсів управління жестами та потенційних застосувань в системах ЛМВ	- словник жестів; налаштування обробки даних; - налаштування параметрів мережі	стратегія спільного використання моделей	інформаційне забезпечення для ЛМВ	методологія розробки і спільного використання візуальних методів розпізнавання жестів рук

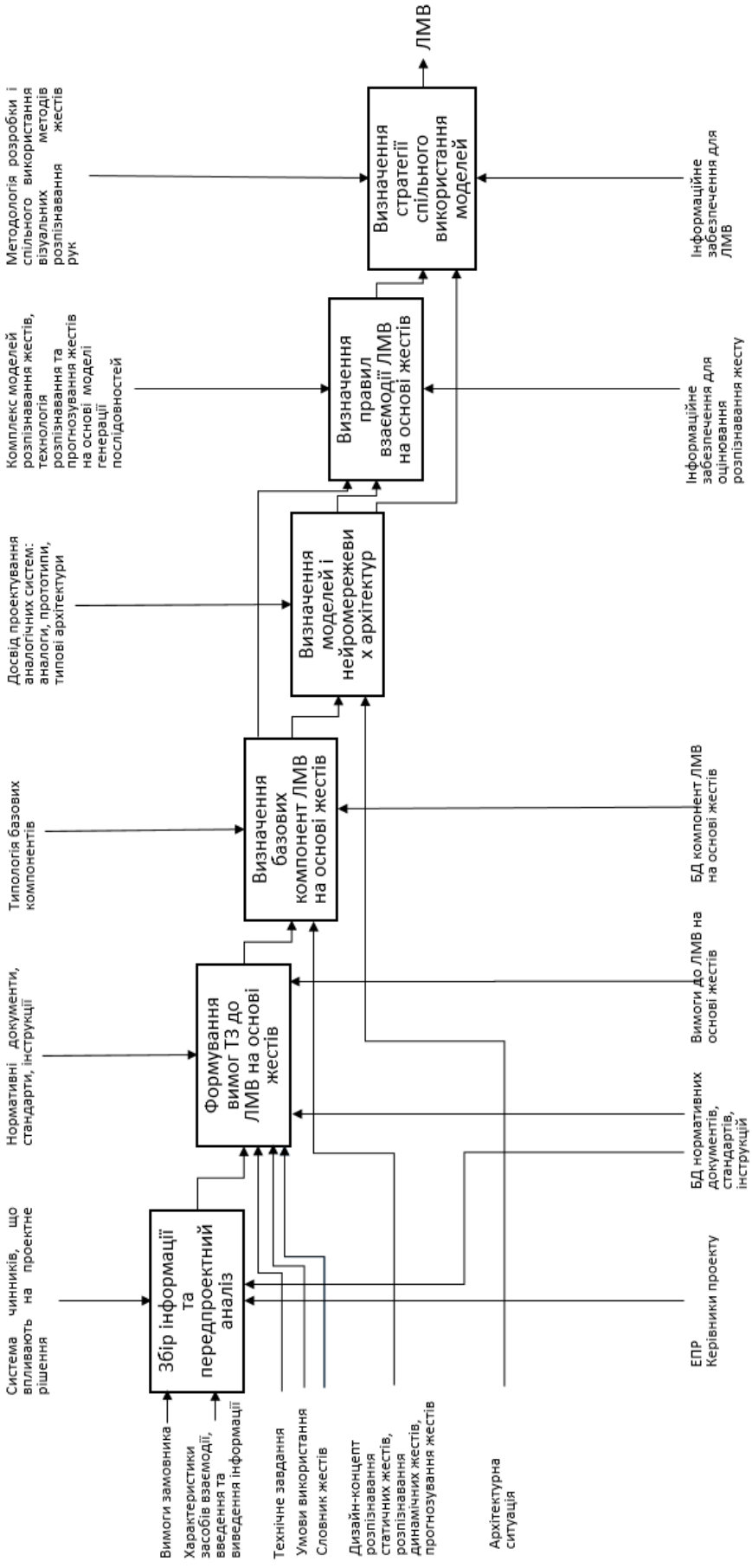


Рисунок 5.2 – Схема розробки інформаційної технології ЛМВ з використанням жестів

5.1.3 Основні етапи розробки інформаційної технології ЛМВ в ОЗ з використанням жестів

Eman 1. Розробка системи ЛМВ починається з визначення потреб, які система повинна задовольнити. Оскільки кінцевою метою даної роботи є реалізація взаємодії хірурга зі спеціалізованим монітором, при безконтактному управлінні переглядом зображень в ОЗ, робота починається з вивчення існуючих способів ЛМВ, які відповідають цим потребам. У випадку використання жестів останні можуть використовуватися, для систем перегляду двох типів медичних зображень, що найчастіше використовуються в ОЗ [3]:

- двовимірні томографічні зображення, які, в основному, відображають зрізи органів, отримані радіо- та КТ-сканами. У цьому випадку взаємодія здійснюється з кожним зображенням окремо в межах одного стеку;
- тривимірні зображення, що є віртуальною реконструкцією органу або групи органів, зібраних в одну структуру. У цьому випадку взаємодія здійснюється одночасно як з однією 3D-моделлю, так і з цілим набором моделей.

Після встановлення випадків використання, необхідно виділили типи взаємодії, які, в свою чергу, дозволяють виділити всі функціональні можливості, на які має бути орієнтована лексика жестів. Далі завдання полягає в тому, щоб пов’язати кожен значущий жест з однією визначеною функціональністю.

Існує дві функціональні можливості, спільні для обох випадків використання: (1) поступальне пересування зображення і (2) масштабування зображення або 3D-моделі. Крім того, кожен випадок використання має особливі типи взаємодії. Так, наприклад, для перегляду двовимірних томографічних зображень, специфічними взаємодіями є: (1) Перегляд стека, коли користувач переглядає складені розділи органів, що складають відповідний випадок використання. Користувач повинен контролювати швидкість перегляду, маючи змогу швидко переглядати великий набір зображень, маючи можливість ізолювати та зупинити певне зображення. (2) Управління контрастністю та

яскравістю: користувач повинен мати можливість збільшувати або зменшувати як контрастність, так і яскравість зображення.

Що стосується випадку управління тривимірними зображеннями, то специфічними взаємодіями є: (1) Обертання і повороти 3D-моделі органів. (2) Можливість увімкнути або вимкнути відображення одне або групу вибраних 3D-моделей органів.

Етап 2. Формування вимог ТЗ до ЛМВ на основі жестів проводиться з використанням технічного завдання, визначених умов використання та сформованого словника жестів. На цьому етапі визначаються вимоги до системи, проводиться розподіл завдань та визначаються критерії успішності реалізації людино-машинної взаємодії.

Етап 3. Визначення базових компонент ЛМВ на основі жестів важливих для конкретного застосування проводиться при врахуванні дизайн-концепту розпізнавання статичних, динамічних жестів та прогнозування жестів. На цьому етапі визначаються основні компоненти та правила людино-машинної взаємодії: визначаються та формуються словник жестів, набір правил відображення, набір інструкцій реагування системи.

Етап 4. Визначення моделей і нейромережевих архітектур, використовуваних для розпізнавання жестів включає визначення складових елементів використовуваної нейронної мережі та її специфікація.

Етап 5. Визначення правил взаємодії ЛМВ на основі жестів відбувається з використанням визначених на етапі 3 словника жестів, набору правил відображення та набору інструкцій, що на виході дозволяють отримати згенеровані інструкції людино-машинної взаємодії. Інструкції генеруються у відповідності до результатів, отриманих з використанням розробленого комплекс моделей розпізнавання жестів, технологія розпізнавання та прогнозування жестів на основі моделі генерації послідовностей.

Етап 6. Визначення стратегії спільного використання моделей, їх налаштування для відповідних інтерфейсів управління жестами та потенційних застосувань в системах ЛМВ. На цьому етапі у відповідності до визначених

словника жестів, параметрів налаштування обробки даних та мережі формується стратегія реалізації людино-машинної взаємодії. Людино-машинна взаємодія реалізується у відповідності до розробленої методології розробки і спільного використання візуальних методів розпізнавання жестів рук.

5.2 Результати тестування моделей та елементів інформаційної технології

Результати тестування та практичної реалізації елементів інформаційної технології, що складається з моделі розпізнавання окремих одиниць жестів, моделі розпізнавання динамічних образів та методу розпізнавання та прогнозування кадрів відеопослідовностей жестів представлено далі.

5.2.1 Особливості реалізації моделі статичного розпізнавання жестів

5.2.1.1 Використовувані дані

Для статичного розпізнавання використовувалися Набір 1, який складався з 12000 зображень, об'єднаних в умовні категорії "one", "two", "zero", "palm", "fist", "letter SH" по 2000 зразків кожен. Зображення для набору даних були отримані покадрово з відеопотоку з веб-камери Logitech HD Webcam C270 з інтервалом між кадрами в 0.05 сек. Дані були зібрані у п’яти осіб при штучному розсіяному світлі.

Процеси отримання даних і розпізнавання проводилися за допомогою однакової веб-камери, встановленої на відстані близько 1 метра до людини, що виконувала жести. Зйомка руки проводилася з різних кутів огляду.

Кадри, зняті послідовно переводилися в чорно-білий одноканальний формат в колірному просторі HSV, після чого, для зниження шуму, пропускалися через фільтр Гаусового розмивання. Зразки кадрів категорій (вони ж класи мережі) представлені на рис. 5.3.

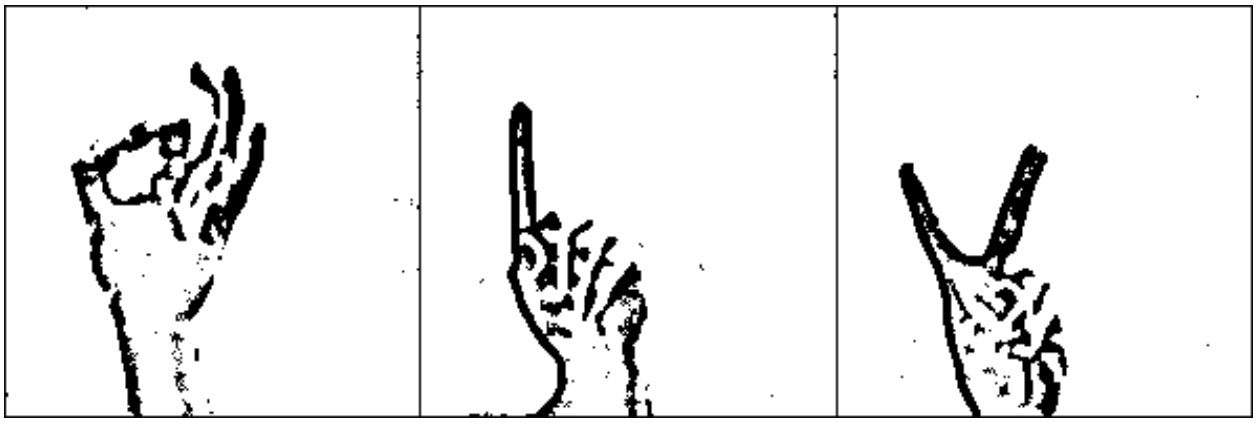


Рисунок 5.3 - Вигляд вхідних даних за трьома категоріями: «zero», «one», «two»

Весь набір даних був розділений на навчальний, валідаційний, і тестовий - призначений для перевірки вже навченої мережі. Вхідний набір шляхом випадкової вибірки даних був розділений на навчальний - 8400 зразка, тестувальний - 1200 зразків та валідаційний в 2400 екземплярів.

Фрагмент послідовних кадрів, використовуваних мережею, представлений на рис. 5.4.

Рисунок 5.4 – Фрагмент кадрів набору для категорії “one”

5.2.1.2 Дизайн експерименту моделі динамічного розпізнавання жестів

Експериментальні дослідження були проведені на комп'ютері з процесором Intel Core i3-7100, 8 Гб оперативної пам'яті, NVIDIA GeForce GTX 1060. Програмне забезпечення написано мовою програмування Python. Були використані бібліотеки кадрів Tensorflow і Keras, які використовують бібліотеки numpy [6], scikit-learn [9], OpenCV [7]. Після серії експериментів в якості моделі для нейронної мережі була обрана згорткова нейронна мережа (Convolutional Neural Network) з чотирнадцяті шарів. Загальна схема мережі у форматі виводу моделі Keras відображена на рис. 5.5.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 32, 198, 198)      320
_________________________________________________________________
activation_1 (Activation)    (None, 32, 198, 198)      0
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 32, 99, 99)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 32, 97, 97)        9248
_________________________________________________________________
activation_2 (Activation)    (None, 32, 97, 97)        0
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 32, 48, 48)        0
_________________________________________________________________
dropout_1 (Dropout)          (None, 32, 48, 48)        0
_________________________________________________________________
flatten_1 (Flatten)          (None, 73728)             0
_________________________________________________________________
dropout_2 (Dropout)          (None, 73728)             0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               9437312
_________________________________________________________________
activation_3 (Activation)    (None, 128)               0
_________________________________________________________________
dropout_3 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_2 (Dense)              (None, 3)                 387
_________________________________________________________________
activation_4 (Activation)    (None, 3)                 0
=================================================================
Total params: 9,447,267
Trainable params: 9,447,267
Non-trainable params: 0
```

Рисунок 5.4 - Загальний вигляд моделі статичного розпізнавання жестів

До згорткових шарів з функцією активації ReLU застосовуються 32 згортальних ядра розміром 3 × 3 з встановленим режимом valid. До отриманих карт ознак застосовувався max-pooling розміром 2 × 2. Також в мережі є стандартний шар flatten, щільний шар на 500 нейронів і softmax-класифікатор з трьома класами. Шари dropout додано для зниження ймовірності перенавчання.

Навчання нейронної мережі тривало 50 епох. За одну епоху оброблялися всі підмножини даних адаптації та валідації. Для перевірки якості використовувалася метрика accuracy, що показує співвідношення кількості правильно передбачених значень до всієї кількості виданих мережею передбачень.

Функцією втрат виступала категоріальна перехресна ентропія, яка вираховувала логарифмічну втрату на кілька представлених класів. В результаті навчання мережі була отримана точність на тестовому підмножині в 98.46%, а значення функції втрат склало - 0.02.

5.2.2 Особливості реалізації моделі динамічного розпізнавання жестів

5.2.2.1 Використовувані дані

Модель динамічного розпізнавання жестів тестувалася з використанням двох наборів даних – Набір 2 (власний) і Набір 3 (вільний набір даних аргентинської мови жестів LSA64 [8]). Для навчання моделі використано дані, що складаються з 30 жестів, кожен з яких має 50 відеопотоків. Для обробки відео кожен потік був розділений на послідовність кадрів у градаціях сірого. В результаті було створено 90000 зображень, що представляють послідовність динамічних жестів, 3000 для кожної категорії.

5.2.2.2 Дизайн експерименту моделі розпізнавання динамічних образів

Експеримент проведено на комп'ютері, оснащеному процесором Intel Core i3-7100, 8 ГБ оперативної пам'яті, NVIDIA GeForce GTX 1060. Програмне забезпечення написане мовою програмування Python. Використано фреймворки

Tensorflow та Keras, бібліотеки numpy, scikit-learn та OpenCV. Розпізнавання руки проводилося з різних кутів огляду.

5.2.3 Реалізація технології розпізнавання та прогнозування жестів

5.2.3.1 Використовувані дані

Модель розпізнавання та прогнозування жестів тестувалася з використанням Набору 3 (вільний набір даних аргентинської мови жестів LSA64 [8]). База даних містить 3200 відео, де 10 непрофесійних випробуваних виконали 5 повторень 64 різних типів знаків.

Запис відео жестів реалізовано з використанням одягу чорного кольору. Жести відтворено стоячи або сидячи, на тлі білої стіни. Для спрощення сегментації рук в межах зображення, випробовувані носили рукавички флуоресцентного кольору. Ці умови істотно спрощують проблему розпізнавання положення руки та виконання її сегментації, а також усувають усі проблеми, пов'язані з варіаціями кольору шкіри, зберігаючи при цьому труднощі розпізнавання форми руки.

Запис відео проведено з застосуванням камери з однаковими технічними параметрами (Sony HDR-CX240). Штатив розміщено на відстані 2 м від стіни на висоті 1,5 м. Відмітки на підлозі використовувались для позначення місця розташування випробовуваних. Розподільна здатність відео становить 1920 на 1080 пікселей, з частотою 60 кадрів в секунду.

В експериментів використано вирізаний варіант відео, який схожий на вихідну версію, але кожне відео було оброблено, так що кадри на початку або в кінці відео без руху в руках були видалені.

Для проведення експерименту використано дані одного жесту – ID: 1, назва: Opaque, що виконується правою рукою. Дані були розділені на 3 компоненти: дані навчання, дані валідації та дані тестування.

5.2.3.2 Дизайн експерименту моделі прогнозування жестів

Тестування запропонованої технології розпізнавання та прогнозування жестів проведено за допомогою Google Collaboration з використанням GPU graphics [2]. В якості контрольної моделі використані ConvLSTM з батч нормалізацією, вихідний код, якої реалізовано Keras [4].

Передобробка даних.

При створенні відео використана рука в кольоровий рукавичці, за рахунок чого процес розпізнавання був спрощений. Передобробка включала в себе виділення контурів руки с допомогою алгоритму виявлення контуру Canny. Алгоритм виявлення контуру Canny складається з 5 кроків: зменшення шуму; розрахунок градієнту; немаксимальне придушення; подвійний поріг; відстеження контуру за допомогою гістерезису.

Навчання та валідація моделі.

Навчання моделі проведено за таких умов. Кількість відеопослідовностей досліджуваного жесту - 50. У кожної відеопослідовності 30 кадрів. Модель навчена на 42-х послідовностях, три послідовності використані для валідації навченої моделі. Розмірність відеокадрів зменшена до 40х40. Таким чином, всі вхідні дані були представлені як масиву розмірністю (50, 30, 40, 40, 1).

Прогнозування відео кадрів.

Одна з двох послідовностей, що не брали участі в навчанні і валідації моделі, використана для прогнозування / генерації кадрів жесту. Після навчання на вхід мережі подається частина кадрів послідовності, яка не брала участі в навчанні і валідації моделі, - в нашому випадку це 10 кадрів. Після 10-го кадру навчена на інших зразках такого ж жесту модель прогнозує / генерує ймовірний рух руки, що виконує жест.

5.3 Оцінка адекватності елементів інформаційної технології

Проведено оцінку адекватності елементів інформаційної технології, що складається з моделі розпізнавання окремих одиниць жестів, моделі

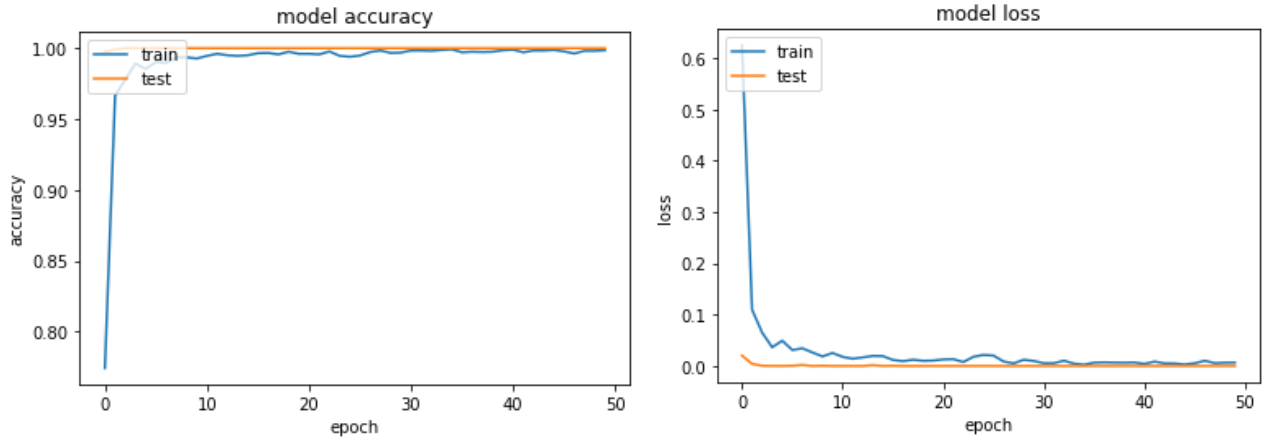
розпізнавання динамічних образів та методу розпізнавання та прогнозування кадрів відеопослідовностей жестів.

5.3.1 Оцінка моделі статичного розпізнавання жестів

Для перевірки якості навчання використовувалася метрика accuracy, тобто, відношення кількості правильно передбачених значень до загальної кількості всіх відповідей. В якості функції втрати $\mathscr{L}$ використано категоріальну перехресну ентропію, тобто вираховувалася логарифмічна втрата на кілька представлених класів. Якщо передбачені моделлю значення дорівнюють q, в той час як справжні значення дорівнюють p, то категоріальна перехресна ентропія буде виглядати як

$$\mathscr{L}(p, q) = -\sum_x \mathrm{p}(x) \log\left(q(x)\right) \tag{4.1}$$

На рис. 5.5 (а) показано зміну показника точність на навчальній і валідаційній вибірках протягом 50-и епох. На рис. 5.5 (б) показана зміна значення функції втрат за той же період.



(а) (б)

Рисунок 5.5 – (а) Оцінки точності моделі статичного розпізнавання для тренувального та тестового наборів, (б) 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 96,12%, а значення функції втрати – 0.02.

5.3.2 Оцінка моделі динамічного розпізнавання жестів

Модель CNN-LSTM, описана в Розділі III, була підготовлена з використанням 1920 відео (з 480 відео перевірки) жестів людської руки із набору даних LSA64. Під час розробки моделі, було проведено кілька експериментів зі зміною значень гіперпараметрів, послідовності шарів та їх кількості.

Експерименти показали, що для вирішення цього завдання достатньо трьох згорткових шарів (CNN), трьох шарів агрегування та двох шарів з довгою короткочасною пам'ятю (LSTM). З метою запобігання перенавчання, коли навчена модель занадто наближається до заданого набору даних і, тим самим, втрачає здатність до узагальнення, у розробленій нейронній мережі використовувалась техніка випадкового виключення нейронів. Це дозволило виключити нейрони певного шару з заздалегідь визначеною ймовірністю на кожному етапі навчання. Це заважає нейронній мережі просто запам’ятовувати правильні відповіді. Окремі вузли виключаються з мережі із встановленою ймовірністю 0,25.

Навчання мережі тривало 25 епох. За одну епоху були оброблені цілі набори даних навчання та перевірки. Навчання проводилося приблизно 15 годин.

Далі була застосована категоріальна перехресна ентропія та розрахована логарифмічна втрата шляхом порівняння прогнозованих та дійсних значень. Відповідно до розрахованих значень, CNN-LSTM використовує оптимізатор Adama [5] із параметрами за замовчуванням $l_г$ = 0,001, β_1 = 0,9, β_2 = 0,999 та = 1e−08.

Як результат, модель показала 98,46% точності тестового набору даних. Значення функції втрат дорівнює 0,001. Робоча модель розпізнає жести рук з

досить високою ймовірністю. На етапі тестування якість модельного навчання перевірялася за допомогою метрики точності, тобто відношення кількості вірно передбачених значень до загальної кількості всіх відповідей.

На рис. 5.6 (а) показана зміна точності в навчальних та валідаційних вибірках протягом 25 епох. На рис. 5.6 (б) показано зміну значення функції втрат за той самий період.

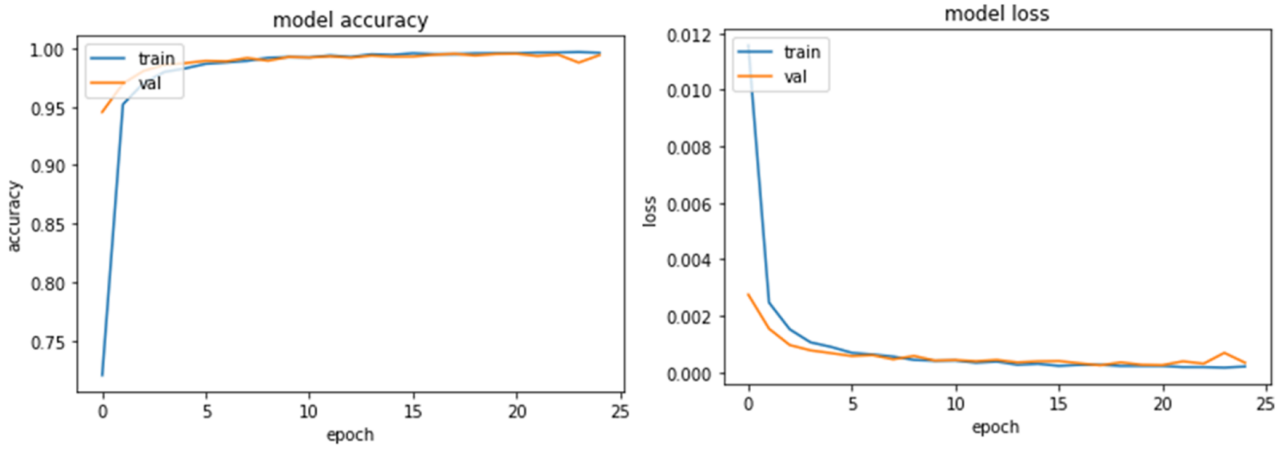


Рисунок 5.6 – (а) Оцінки точності моделі динамічного розпізнавання жестів для тренувального та тестового наборів, (б) функції втрат для для тренувального та тестового наборів

Як показано на рис. 5.6 (б), значення вказують на те, що значення функції втрат швидко зменшуються, і, починаючи з 4-ї епохи, залишається незмінно низьким.

Ефективність класифікаційної моделі показана на рис. 5.7, де представлена нормалізована матриця невідповідності моделі. Рядки відповідають передбачуваній категорії, а стовпці відповідають фактичній категорії.

З рис. 5.7 видно, що для Набору 3, кількість неправильно класифікованих жестів є досить низькою. Найбільш складним випадком для класифікації, були жести women і son з рівнем перекриття 0,02.

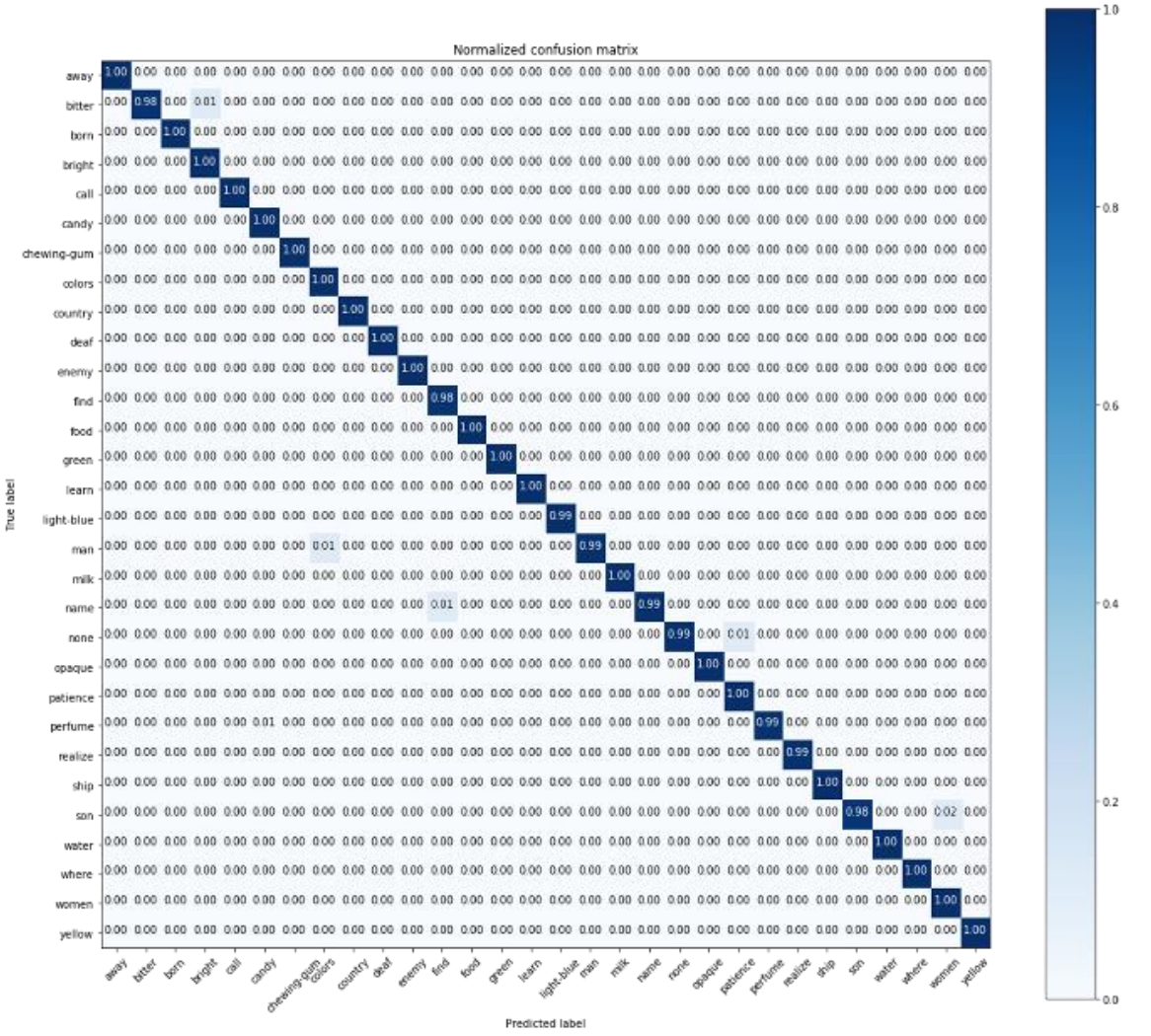


Рисунок 5.7 – Нормалізована матриця невідповідності

5.3.3 Оцінка технології прогнозування відеопослідовностей жестів

Для оцінки досягнутого поліпшення якості прогнозування, ми порівнюємо результати, отримані при використанні запропонованого підходу, з результатами прогнозування, отриманими при використанні ConvLSTM з пакетною нормалізацією.

5.3.3.1 Оцінка результатів розпізнавання

Кількісна оцінка отриманих результатів проведена з використанням метрик точності та бінарної перехресної ентропії, що представляє функцію втрат.

На рисунку 5.8 представлено кількісну оцінку точності моделі бінарної крос-ентропії, що представляє функцію втрат, для кожної прогнозованої

послідовності для навчання та валідації з використанням запропонованої технології.

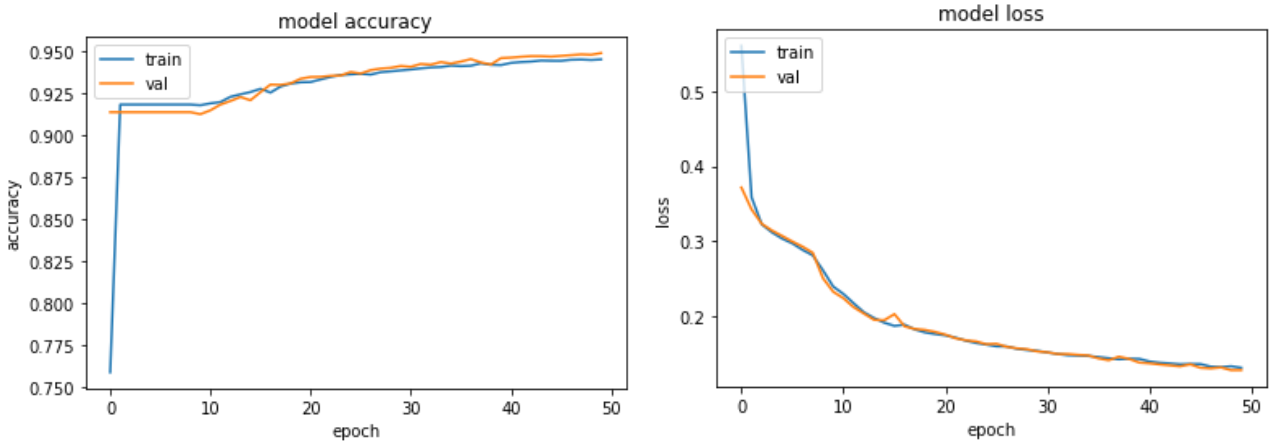


Рисунок 5.8 – Оцінки точності та бінарної крос-ентропії отримані

На рисунку 5.9 представлено кількісне оцінювання точності моделі та бінарної перехресної ентропії, що представляє функцію втрат для навчання та валідації з використанням ConvLSTM із пакетною нормалізацією.

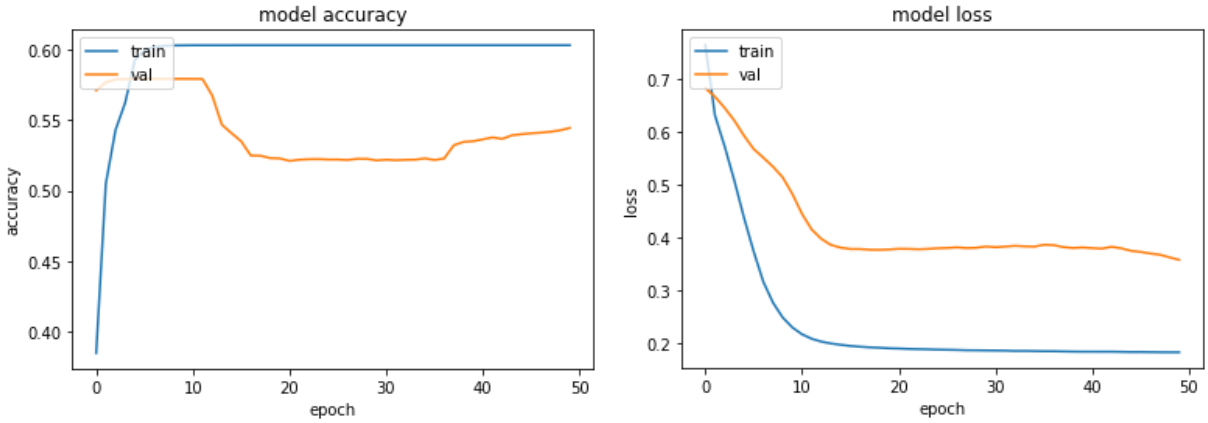


Рисунок 5.9 – Оцінки точності та бінарної крос-ентропії отримані з використанням ConvLSTM із батч нормалізацією

Як видно при порівнянні результатів, що представлено на рис. 5.8 та 5.9, запропонований метод на етапі навчання та валідації надає точність набагато вище, ніж при використанні базової моделі ConvLSTM з пакетною

нормалізацією. Точність при використанні ConvLSTM з пакетною нормалізацією не вища за 0,6, тоді як при валідації з використанням запропонованого методу точність доходить до 0,9. Крім того, можна замінити те, що метрика Binary Cross-Entropy / Log Loss, яка є помилкою моделей при навчанні та валідації при використанні запланованого підходу менше, ніж у ConvLSTM з нормалізацією партії. При використанні ConvLSTM із пакетною нормалізацією не опускається нижче 0,35, тоді як при валідації з використанням запропонованого підходу виявляється помилка менше 0,2.

5.3.3.2 Якісний результат прогнозування

Якісний результат багатоетапного прогнозування мови жестів на даних жеста - ID: 1, назва: Opaque. 10 послідовних зображень у першому рядку на рис. 5.10 - це вхідні дані нейронної мережі, а наступні чотири рядки – істині кадри відео послідовності для ConvLSTM з пакетною нормалізацією, прогнозовані результати ConvLSTM з пакетною нормалізацією, істині кадри відео послідовності для кадрів, прогнозованих з використанням запропонованого методу, та прогнозовані результати запропонованого методу, відповідно (рис. 5.10).

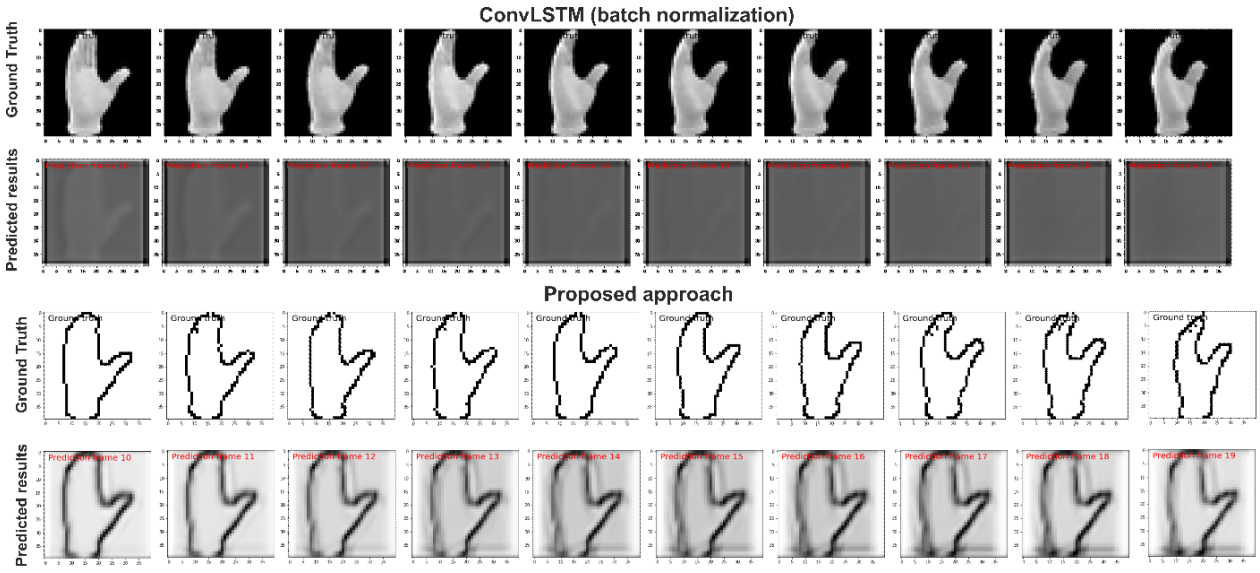


Рисунок 5.10 – Якісне порівняння результату

При якісному порівнянні результатів прогнозування, отриманих при використанні базових моделей та запропонованого підходу, помітно що більш точне прогнозування жестів рук отримано при використанні запропонованого підходу, що підтверджує кількісну оцінку результатів прогнозування, отриманих за допомогою метрики нормалізованої перехресної кореляції.

Візуальна якісна оцінка результатів прогнозування забезпечує перевагу запропонованого підходу. Це пояснюються тим, що при використанні контурування жестів ми отримуємо більш значущу інформацію для прогнозування / генерування жестів відео кадрів.

5.3.3.3 Кількісний результат якості прогнозування

Оцінка 10 прогнозованих кадрів проведена при їх порівнянні з реальними кадрами з використанням евклідової, манхеттенської відстані та нормалізованої перехресної кореляції. Евклідова і манхеттенська відстані вимірюють різницю між прогнозованими та реальними кадрами. При аналізі отриманих результатів необхідно брати до уваги метрики, які є бажаними для даних високих розмірностей, якими э дані послідовності відеокадрів. Для даної задачі з великим значенням розмірності даних може бути кращим використовувати нижчі значення. Це означає, що метрика манхеттенської відстані є найбільш переважною для даних великої розмірності [1]. Таким чином, манхеттенська відстань є найбільш прийнятним параметром для оцінки ніж евклідова метрика відстані, оскільки розмір даних збільшується.

Оцінено показник взаємної кореляції між невеликою частиною поточного прогнозованого кадра та його локальною окремістю у попередньому реальному кадрі. Передбачено, що елементи руху, наявні на всій сцені (кадру), ефективно апроксимуються сусідніми просторовими блоками більш низького рівня, використовуючи невеликі локальні орієнтації у часовому вимірі. Це пов'язано з тим фактом, що, якщо відео не містить значущого тремтіння або непередбачених випадкових подій, таких як зміна сцен, функції руху залишаються плавними протягом часу.

Результати для 10 прогнозованих кадрів, отриманих із використанням ConvLSTM з пакетною нормалізацією, у порівнянні з реальними кадрами, представлені в таблиці 5.3.

Таблиця 5.3 - Результати оцінки прогнозування з використанням евклідової, манхеттенської відстаней та нормалізованої перехресної кореляції для ConvLSTM з пакетною нормалізацією

Метрика	Кадр 10	Кадр 11	Кадр 12	Кадр 13	Кадр 14	Кадр 15
Евклідова відстань	0.00061	0.00061	0.00059	0.00059	0.00060	0.00060
Манхетенська відстань	0.13987	0.13877	0.13886	0.14032	0.14173	0.14172
Нормалізована крос-кореляція	0.09638	0.06622	0.03272	0.01243	-0.00358	-0.01721

Метрика	Кадр 16	Кадр 17	Кадр 18	Кадр 19	Середнє знач. метрики
Евклідова відстань	0.00060	0.00061	0.00061	0.00060	0.000602
Манхетенська відстань	0.14348	0.14418	0.14448	0.14305	0.141646
Нормалізована крос-кореляція	-0.03392	-0.03822	-0.04318	-0.06131	0.001033

Результати для 10 прогнозованих кадрів, отриманих з використанням запропонованого підходу, в порівнянні з реальними кадрами, представлені в таблиці 5.4.

З табл. 5.3, 5.4 можна зробити висновок, що евклідова і манхеттенська відстані між прогнозованими і реальними даними з кожним кадром збільшується, що вказує на все більшу відміну прогнозованого кадру від реального.

Таблиця 5.4 - Результати оцінки прогнозування з використанням евклідової, манхеттенської відстаней і нормалізованої перехресної кореляції для запропонованого підходу

Метрика	Кадр 10	Кадр 11	Кадр 12	Кадр 13	Кадр 14	Кадр 15
Евклідова відстань	0.00071	0.00073	0.00076	0.00081	0.00080	0.00080
Манхеттенська відстань	0.11838	0.11946	0.11958	0.12236	0.11803	0.11503
Нормалізована крос-кореляція	0.61510	0.51900	0.42948	0.35369	0.28873	0.23654

Метрика	Кадр 16	Кадр 17	Кадр 18	Кадр 19	Середнє знач. метрики
Евклідова відстань	0.00080	0.00081	0.00081	0.00081	0.00078
Манхеттенська відстань	0.11152	0.10743	0.10598	0.10363	0.11414
Нормалізована крос-кореляція	0.22801	0.17827	0.14165	0.14899	0,313946

На цей же факт вказує значення нормалізованої перехресної кореляції, що зменшується з кожним кадром. Також, можна помітити, що твердження [1] підтвердилося.На даних відеопослідовностей високих розмірностей значення евклідової відстані для кількох послідовних кадрів повторюється, що обґрунтовує переважне використання результатів манхеттенської відстані для оцінки результатів прогнозування відео послідовностей.

Як можна побачити в табл. 5.3, 5.4 результати прогнозування запропонованої моделі по метриках манхеттенської відстані і нормалізованої перехресної кореляції краще, ніж у ConvLSTM з пакетною нормалізацією. Метрика евклідової відстані навпаки, показує велику схожість між точками даних прогнозованого і реального кадру. Уже на першому прогнозованому кадрі манхеттенська відстань між реальним кадром і прогнозованим на 0.02149 менше

у запропонованого методу, що говорить про меншу відмінність прогнозованого і реального кадру. Нормалізована перехресна кореляція на першому кадрі прогнозованому за допомогою запропонованого методу становить 0.61510, що говорить про високу залежність даних. Нормалізована перехресна кореляція на першому прогнозованому кадрі між реальним кадром і прогнозованим на 0.51872 більше у запропонованого методу, що говорить про більшу залежність точок даних прогнозованого і реального кадру. Евклідова відстань показує меншу різницю при використанні ConvLSTM з пакетною нормалізацією, що пояснюється високою розмірністю даних.

Середні оцінки для 10 прогнозованих і реальних кадрів представлені в табл.5.5.

Таблиця 5.5 - Середні оцінки прогнозування з використанням евклидова, манхеттенського відстаней і нормалізованої крос-кореляції для запропонованого методу і ConvLSTM з пакетною нормалізацією

Метрика	ConvLSTM з пакетною нормалізацією	Запропонований метод
Евклідова відстань	0.000602	0.00078
Манхеттенська відстань	0.141646	0.11414
Нормалізована крос-кореляція	0.001033	0,313946

Середнє значення манхеттенської відстані на десяти прогнозованих кадрах при використанні запропонованого методу і ConvLSTM з пакетною нормалізацією становить 0.11414 та 0.141646 відповідно. Запропонований метод показав середню відмінність кадрів на 0.027506 менше, ніж у моделі ConvLSTM з пакетної нормалізацією. Це говорить про більшу схожість кадрів, отриманих з використанням запропонованого методу. Середня нормалізована крос-кореляція на 10 прогнозованих кадрах більше у запропонованого методу на 0.312913, що говорить про більшу схожість прогнозованого і реального кадру. Евклідова відстань показує меншу різницю кадрів у моделі ConvLSTM з пакетної

нормалізацією, що пояснюється високою розмірністю даних і некоректністю використання в цьому випадку цієї метрики.

5.3.4 Результат застосування моделі скінченного автомату з використанням моделей розпізнавання та прогнозування жестів

Модель скінченного автомату реалізується з використанням прогнозованих кадрів, що отримуються з використанням методу розпізнавання та прогнозування кадрів відеопослідовностей жестів.

Початок жесту, який використовується та на підставі якого здійснюється прогнозування, складає 10 кадрів. На підставі цих 10 кадрів отримуються прогнозовані наступні кадри, які поступають на вхід моделі скінченного автомату.

На підставі записаного набору жестів, визначена середня тривалість жесту, що складає 1,5 секунд. Відео має наступні характеристики: розподільна здатність дорівнює 720x720, 30 кадрів на секунду.

Таким чином, розрахунок часу, що використовується на обробку відео представлений наступним чином.

Без використання прогнозування (використовуємо середню тривалість відео одного жесту) час розпізнавання та відповідної реакції системи в режимі реального часу залежить від тривалості жесту, та, в середньому, при ідеальних умовах розпізнавання, має затримку в 1,5 секунди, що дорівнюється 45 кадрам.

З використанням методу прогнозування час розпізнавання та відповідної реакції системи в режимі реального часу не залежить від тривалості жесту, та, в середньому, при ідеальних умовах розпізнавання, має затримку в 0,75 секунди, або 25 кадрів.

Таким чином, удосконалено технологію безконтактної взаємодії системою управління медичними зображеннями в операційній з використанням жестів рук та зменшено затримку відклику системи на 0,75 секунди, що в 2 рази краще за час відклику без використання запропонованої інформаційної технології.

Разом з тим, варто зазначити, що згідно [10], в режимі реального часу система повинна мати можливість завершити конвеєр обробки не більше ніж 250 мілісекунд, оскільки середній час реакції на зорові подразники коливається від 250 до 350 мілісекунд. Більший діапазон призводить до візуально помітного відставання. Отже, в майбутньому необхідні додаткові експерименти для отримання найкращого результату, що гарантуватиме ЛМВ без замітного відставання.

Очевидно, що при різних видах хірургічних процедур існуватимуть різні обмеження щодо того, як і коли можливості маніпуляцій із зображеннями можуть поєднуватися при використанні хірургічного інструменту. Це вимагатиме ретельного розгляду способу здійснення введення, особливо у випадках, коли обидві руки можуть тримати інструменти. У таких випадках може бути можливо використовувати голосові команди для різних маніпуляцій із вільними руками (за умови, що вони відповідають дискретним властивостям голосових команд) або поєднувати голос з іншими типами введення, такими як ножні педалі, введення погляду або рух головою.

Висновки до п’ятого розділу

1. Удосконалено структурну модель інформаційної технології ЛМВ з використанням жестів, шляхом визначення основних етапів та інформаційних потоків створення та інтеграції моделей глибокого навчання, яка забезпечує прийняття рішень щодо застосування розроблених методів, засобів і технологій.
2. Проведено тестування розробленої системи розпізнавання статичних и дінамічних жестів на різних вибірках, як публічно доступних, так і створених самостійно. Проведена тестування технології прогнозування кадрів у відеопотоці по декількох метриках, наведені результати і складено відповідні таблиці.
3. За результатами тестування зроблено висновок, що реалізована система має високу ефективність по точності розпізнавання і швидкості обробки, що

дозволяє працювати з відеопотоком в режимі реального часу. Запропонована система стійка до всіх видів перетворення об'єктів, включаючи повороти, масштабування, переміщення.

4. Запропонована технологія безконтактної взаємодії для управління медичними зображеннями з використанням жестів рук дозволила зменшити затримку відклику системи на 0,75 секунди, що в 2 рази краще за час відклику без використання запропонованої інформаційної технології.

Результати розділу опубліковано в роботах автора [3, 4, 6, 9, 11, 12, 15, 16] (Додаток А)

Література до п’ятого розділу

1. Aggarwal, C.C., Hinneburg A., Keim D.A. (2001) On the surprising behavior of distance metrics in high dimensional space. In International conference on database theory, Springer, Berlin, Heidelberg, pp. 420-434.

2. Google Collaboratory Режим доступу: https://colab.research.google.com /notebooks/intro.ipynb (21.09.2020).

3. Hurstel A., Bechmann D. (2019) Approach for Intuitive and Touchless Interaction in the Operating Room, J 2, no. 1: pp. 50-64. https://doi.org/10.3390/j2010005

4. Keras Режим доступу: https://keras.io/examples/conv_lstm/ (21.09.2020).

5. Kingma D.P., Ba J. (2014) Adam: A method for stochastic optimization, arXiv preprint arXiv: 1412.6980.

6. Numpy. Режим доступу: https://numpy.org/ (21.09.2020)

7. OpenCV. Режим доступу: https://opencv.org/ (21.09.2020)

8. Ronchetti F., Quiroga F., Estrebou C.A., Lanzarini L.C., Rosete A. (2016). LSA64: An Argentinian Sign Language Dataset. Web

9. Scikit-learn. Режим доступу: https://scikit-learn.org (21.09.2020)

10. Shelton J., Kumar G. (2010) Comparison between Auditory and Visual Simple Reaction Times. Neuroscience & Medicine vol. 1, pp. 30-32. doi:10.4236/nm.2010.11004.

11. Wan J., Zhao Y., Zhou S., Guyon I., Escalera S., Li S.Z. (2016) Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition. In Proceedings of the 2006 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 56-64.

ВИСНОВКИ

1. У дисертації поставлено і розв’язано актуальну науково-прикладну задачу розроблення моделей та методу інформаційної технології для систем комп’ютерного зору як основи людино-машинної взаємодії з використанням жестів рук.

Отримано нові теоретичні та практичні результати:

1. На основі проведеного аналізу існуючих підходів і методів ЛМВ встановлено, що існує об’єктивне протиріччя між зростаючою кількістю впроваджень високотехнологічних систем комп’ютерного зору і рівнем адаптації цих технологій в системах ЛМВ за допомогою жестів. Головною вимогою візуального розпізнавання жестів до їх використання в інтерфейсах ЛМВ є безперервне розпізнавання в режимі реального часу з мінімальною затримкою на виході розпізнавання, але досягти такої продуктивності системи надзвичайно складно з наступних причин: (1) початкова і кінцева точка жесту невідомі; (2) захоплений кадр має негайно пройти попередню обробку, після чого всі елементи повинні бути витягнуті та використані для класифікації, перш ніж буде захоплено наступний кадр; (3) необхідно встановити спеціальні умови щодо освітлення, фону, розташування і людини і камери.

2. Попри значні дослідницькі досягнення в сфері безконтактного управління жестами, у багатьох процесах розробки все ще бракує врахування ергономічних вимог, що може мати наслідки для здоров'я та досвіду користувачів при тривалому використанні.

3. З метою забезпечення ефективного процесу проектування систем ЛМВ у дисертації пропонується нова методологія розробки моделей розпізнавання статичних та динамічних жестів рук за допомогою CNN та CNN+LSTM, що дозволило отримати дорожню карту з розробки моделей розпізнавання жестів, способів їх налаштування для різних інтерфейсів управління жестами та потенційних застосувань в системах ЛМВ.

4. Проведено розробку моделей комп’ютерного зору на основі глибоких

нейронних мереж, що забезпечують відстеження та розпізнавання статичних і динамічних жестів рук, здатних функціонувати в режимі реального часу, інваріантних до афінних перетворень і зміни освітлення. Зокрема удосконалено модель статичного розпізнавання жестів, яку побудовано згідно пропонованої методології на основі згорткової нейронної мережі, шляхом штучного збільшення даних і використання контурів. Завдяки використанню контурів, модель є стійкою до відносно широких кутів обертання рук і незалежною від освітлення. Штучне збільшення даних дозволило покращити точність розпізнавання жестів на 4% по відношенню з моделлю без аугментації.

5. Дістала подальшого розвитку технологія розпізнавання та прогнозування жестів на основі моделі генерації послідовностей з використанням ConvLSTM шляхом введення контурів на етапі попередньої обробки, реалізованих у вигляді фільтру, що дозволило фіксувати часові залежності та зменшити шум вхідного сигналу. В базовій моделі була проведена заміна 3 шарів batch normalization на 3 шари dropout regularization що дозволило знизити перенавчання та мінімізувати помилки передбачення при прогнозуванні жестів.

6. Проведено удосконалення моделі скінченного автомату для безконтактного управління переглядом медичних зображень шляхом її адаптації до використання даних прогнозованих кадрів відеопослідовностей з моделі ConvLSTM, що дозволило зменшити час відгуку системи;

7. Створено новий набір даних і виконано концептуальну розробку інтуїтивного словникового запасу динамічних жестів, що дозволило реалізувати ефективну безконтактну інтерактивну систему, адаптовану до особливостей хірургічного контексту.

8. Удосконалено структурну модель інформаційної технології ЛМВ з використанням жестів шляхом визначення основних інформаційних потоків створення та інтеграції моделей глибокого навчання, яка забезпечує прийняття рішень щодо застосування розроблених методів, засобів і технологій.

9. Усі теоретичні положення дисертації доведено до відповідних

інженерних рішень із застосуванням запропонованої інформаційної технології. Практичні результати дисертаційної роботи апробовано та впроваджено в медичному центрі “Твоя лікарня” при розробці проекту клінічної системи безконтактного перегляду медичних зображень у м. Сєвєродонецьк, Луганській області.

10. Матеріали дисертації використані і впроваджені в навчальний процес на кафедрі комп'ютерних наук та інженерії Східноукраїнського національного університету ім. В. Даля.

Тим самим доведено, що поставлені завдання виконано, мету роботи досягнуто.

ДОДАТОК А
СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

Праці, які відображають основні наукові результати дисертації

1. Skarga-Bandurova I. Siriak R., Bilodorodova T., Cuzzolin F., Bawa V.S., Mohamed M.I., Charles R.D. “Surgical Hand Gesture Prediction for the Operating Room”, Studies in Health Technology and Informatics (Eds. Bernd Blobel, Lenka Lhotska, Peter Pharow, Filipe Sousa). 2020, Vol. 273, pp. 97-103. (Індексується в SCOPUS).

2. Сіряк Р.В., Скарга-Бандурова І.С., Шумова Л.О. “Особливості технології обробки даних для розпізнавання жестів”, *Вісник Національного технічного університету "ХПІ". Збірник наукових праць. Серія: Інформатика та моделювання*. Харків: НТУ "ХПІ". 2019, №.13(1338). С. 112-122.

3. Білобородова Т.О., Сіряк Р.В., Скарга-Бандурова І.С., Давіденко М.О., Сіроштан І.В., Приймак С.О. “Розпізнавання взіємодії інструментів з тканиною на медичних відеозображеннях”, *Наукові вісті Далівського університету : електронне наукове фахове видання.* 2019, № 17. Режим доступу: http://nvdu.snu.edu.ua/wp-content/uploads/2020/03/2019_17_5.pdf

4. Сіряк Р.В., Скарга-Бандурова І.С., Білобородова Т.О. Towards an empirical hyper parameters optimization in CNN”, *Вісник Східноукраїнського національного університету ім. В. Даля*. Сєвєродонецьк: СНУ ім. В. Даля. 2019, № 5(253). С. 87-91.

5. Сіряк Р.В., Скарга-Бандурова І.С. “Модель обробки потокових даних для розпізнавання окремих одиниць жестової мови”, *Вісник Національного технічного університету "ХПІ". Збірник наукових праць. Серія: Інформатика та моделювання*. Харків: НТУ "ХПІ". 2018, № 42(1318). С. 73-81.

6. Болтов Є.В., Сіряк Р.В., Щербаков Є.В., Лифар О.К. “Тестування продуктивності операційних систем реального часу для комп’ютерного зору і обробки зображень на одноплатних комп’ютерах”, *Вісник Східноукраїнського*

національного університету ім. В. Даля, Сєвєродонецьк: СНУ ім. В. Даля. 2018, № 6(247). С. 23-26.

7. Сіряк Р.В. “Технологии идентификации и распознавания жестов”, *Вісник Східноукраїнського національного університету ім. В. Даля*, Сєвєродонецьк: СНУ ім. В. Даля. 2017, № 8(238). С.79-85.

Наукові праці, які засвідчують апробацію матеріалів дисертації

8. Siriak R., Skarga- Bandurova I., Boltov Y. “Deep Convolutional Network with Long Short-Term Memory Layers for Dynamic Gesture Recognition”, *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2019 10th IEEE International Conference,* 2019. Vol. 1, pp. 158-162. (Індексується в SCOPUS).

9. Boltov Y., Hrushka M., Kotsiuba I., Skarga-Bandurova I., Krivoulya G., Siriak R. “Performance Evaluation of Real-Time System for Vision-Based Navigation of Small Autonomous Mobile Robots”, *2019 IEEE 10th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, UK, Leeds, June 5-7, 2019. pp. 218-222. (Індексується в SCOPUS).

10. Сіряк Р.В., Скарга-Бандурова І.С. Система розпізнавання динамічних жестів за допомогою мережі CNN-LSTM. *Комп’ютерні інтелектуальні системи та мережі. Матеріали XI Всеукраїнської науково практичної WEB конференції аспірантів, студентів та молодих вчених* (21-23 березня 2019 р.). – Кривий Ріг: ДВНЗ «Криворізький національний університет», 2019. – С. 94-95.

11. Siriak R., Skarga-Bandurova I., Biloborodova T. Hyperparameters Optimization in CNN for Image Recognition. *Theoretical and Applied Computer Science and Information Technology: Proceedings of the III International Conference TACSIT-2019,* May 8-9 2019. Severodonetsk: Volodymyr Dahl East Ukrainian National University, pp. 21-22.

12. Сіряк Р.В., Скарга-Бандурова І.С. Аналіз ефективності адаптивних методів оптимізації гіперпараметрів згорткових нейронних мереж. *Проблеми*

інформатики і моделювання. Тезиси дев'ятнадцятої міжнар. наук.-техн. конф. – Харків: НТУ "ХПІ", 2019. С. 76.

13. Сіряк Р.В., Скарга-Бандурова І.С. Використання конволюційної нейронної мережі для розпізнавання динамічних образів. *Сучасні технології в освіті та науці: Матеріали міжнар. наук.-практ. конф*.; 19 – 22 лютого 2018 р., м. Сєвєродонецьк / гол. ред. О.І. Рязанцев – Сєвєродонецьк: вид-во СНУ ім. В. Даля, 2018. С. 45-47.

14. Сіряк Р.В., Скарга-Бандурова І.С. Вибір архітектури глибинного навчання для системи розпізнавання жестів рук. *Інформаційні технології: наука, техніка, технологія, освіта, здоров'я: тези доповідей XXVI міжнародної науково-практичної конференції MicroCAD-2018*, 16-18 травня 2018р.: у 4 ч. Ч. IV. / за ред. проф. Сокола Є.І. – Харків: НТУ «ХПІ». С. 203.

15. Сіряк Р.В. Средства создания отчетной медицинской документации в медицинской информационной системе. *Збірник тез доповідей міжнародної конференції студентів, аспірантів та молодих вчених "Технологія-2012"*, Сєвєродонецьк: Технол. ін-т Східноукр. Нац. ун-ту ім. В. Даля (м. Сєвєродонецьк), 2012. С. 94-95.

16. Сіряк Р.В., Висоцький Д.В. Система розпізнавання жестів рук через реалізацію моделі нейронної мережі. *ІТ-Ідея – 2018: збірник науково-практичних праць*. Сєвєродонецьк : Вид-во Східноукр. ун-ту ім. В. Даля, 2018. С. 80-81.

ДОДАТОК Б
АКТИ ВПРОВАДЖЕННЯ

“ЗАТВЕРДЖУЮ”
Директор медичного центру
“Твоя лікарня”

Юркова І.А.

2019 р.

А К Т
про впровадження в роботу медичного центру “Твоя лікарня”
результатів дисертаційного дослідження
Сіряка Ростислава Вікторовича,
представленої до захисту на здобуття вченого ступеня
кандидата технічних наук

Науково-медична комісія у складі:
- головний лікар Юрков А.Н.,
- заст. головного лікаря Прищепа О.Л.

склала цей акт про те, що результати дисертаційної роботи Сіряка Р.В.:

1. Методологія розробки і спільного використання візуальних методів розпізнавання жестів рук, яка дозволяє створювати і досліджувати моделі глибокого навчання для розпізнавання статичних та динамічних жестів, шляхом систематизації додаткових процедур збільшення даних, підготовки і попередньої обробки даних разом з процедурою використання моделей, важливих для конкретного застосування;
2. Модель статичного розпізнавання жестів, побудованої згідно з пропонованою методологією на основі згорткової нейронної мережі;
3. Технологія розпізнавання та прогнозування динамічних жестів на основі попередньої фільтрації жестів та генерації послідовностей з використанням моделі ConvLSTM.
4. Інформаційна технологія людино-машинної взаємодії з використанням жестів для управління переглядом двовимірних топографічних зображень на хірургічному моніторі на основі інтеграції технологій глибокого навчання та кінцевих автоматів
використані при розробці робочого прототипу гібридної операційної зали.

Застосування запропонованих в роботі методів, моделей та інформаційної технології дозволяють зробити наступні висновки:

1. В процесі впровадження інформаційної технології підтверджено достовірність і адекватність запропонованих математичних моделей, методу та алгоритмів.
2. Розроблена інформаційна технологія дозволила розширити можливості безконтактного керування зображеннями, використовуючи динамічні жести двома руками та контекстуальні знання для поліпшення навігаційних характеристик
3. Тестове впровадження показало, що система безперерв ного розпізнавання жестів успішно виявляє жести в 94,25% випадків із надійністю 89,98%.
4. Експериментальні результати показали суттєві покращення часу виконання завдань завдяки ефекту інтеграції контексту.

Головний лікар Юрков А.Н.

Заст. гол. лікаря Прищепа О.Л.

ЗАТВЕРДЖУЮ
Проректор з наукової роботи Східноукраїнського національного університету імені Володимира Даля
_____ О.Б. Целіщев
«_4_» _____ 2021 р.

АКТ
про впровадження в навчальний процес
Східноукраїнського національного університету ім. Володимира Даля
результатів дисертаційного дослідження здобувача кафедри
комп'ютерних наук та інженерії

Сіряка Ростислава Вікторовича

"Моделі та метод інформаційної технології людино-машинної взаємодії з використанням жестів"

Цим актом підтверджується, що результати досліджень, що проводились у межах наукового напряму "Інформаційні технології в промисловості, екології, медицині" за науково-дослідними роботами "Система управління медичною інформацією" (ДР № 0111U001748), "Дослідження методів аналізу даних в медицині" (ДР № 0116U008700), проектом Європейського Союзу ERASMUS+ ALIOT 573818-EPP-1-2016-1-UK-EPPKA2-CBHE-JP "Internet of Things: Emerging Curriculum for Industry and Human Applications"; тематичними планами науково-дослідних робіт Східноукраїнського національного університету ім. В. Даля протягом 2012-2021 рр. впроваджені на кафедрі комп'ютерних наук та інженерії в процес підготовки здобувачів вищої освіти за спеціальністю 122 – комп'ютерні науки.

Впровадження результатів дисертаційної роботи полягає в їх використанні при викладанні навчальних дисциплін як окремих розділів лекційних курсів, так і в циклах практичних і лабораторних робіт.

Отримані матеріали досліджень дозволили розробити практикуми з наступних дисциплін:

1. "Машинне навчання" розроблено та впроваджено практичні роботи за темами "Нейронні мережі та глибоке навчання", "Згорткові нейронні мережі".

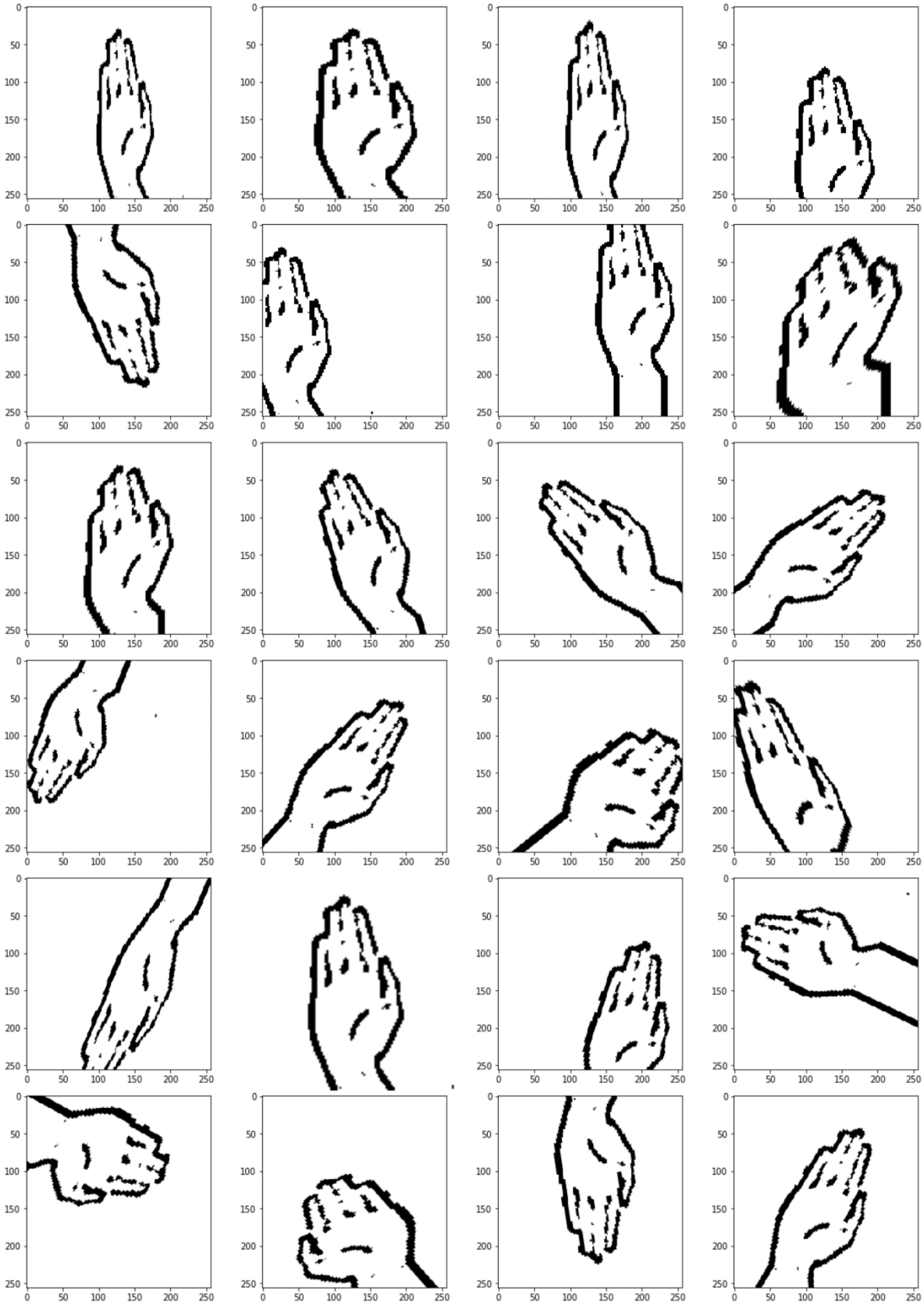
2. "Вступ до інтерактивного проектування" введені матеріали для практичного засвоєння "Засоби людино-машинної взаємодії", "Безконтактні інтерфейси та управління засобами візуалізації даних з використанням жестів".

Декан факультету інформаційних технологій та електроніки, к.т.н., доц. С.О. Митрохін

Завідувач кафедри комп'ютерних наук та інженерії, д.т.н., проф. О.І. Рязанцев

ДОДАТОК В

ВИХОДИ РІЗНИХ МЕТОДІВ ЗБІЛЬШЕННЯ ДАНИХ, ЩО ВИКОРИСТОВУЮТЬСЯ У НАБОРІ ДАНИХ 1, ЗАСТОСОВАНІ ДО ВИБІРКОВОГО КАДРУ



ДОДАТОК Г

ПРОГРАМНА РЕАЛІЗАЦІЯ МОДЕЛІ ПРОГНОЗУВАННЯ ЖЕСТІВ

```
from keras.models import Sequential
from keras.layers.convolutional import Conv3D
from keras.layers import Dropout
from keras.layers.convolutional_recurrent import ConvLSTM2D
from keras.layers.normalization import BatchNormalization
import numpy as np
import pylab as plt
import pickle

import os
from os.path import join, exists
from PIL import Image

imgRows = 40
imgCols = 40
imgDepth = 30

seq = Sequential()
seq.add(ConvLSTM2D(filters=30, kernel_size=(3, 3),
                        input_shape=(None, imgRows, imgCols, 1),
                        padding='same', return_sequences=True))
seq.add(Dropout(0.2))

seq.add(ConvLSTM2D(filters=30, kernel_size=(3, 3),
                        padding='same', return_sequences=True))
seq.add(Dropout(0.2))

seq.add(ConvLSTM2D(filters=30, kernel_size=(3, 3),
                        padding='same', return_sequences=True))
seq.add(Dropout(0.2))

seq.add(Conv3D(filters=1, kernel_size=(3, 3, 3),
                    activation='sigmoid',
                    padding='same', data_format='channels_last'))
seq.compile(loss='binary_crossentropy',optimizer='adam',
metrics=['accuracy'])


from keras.utils.vis_utils import plot_model
import matplotlib.image as mpimg
# Save our model diagrams to this path
model_diagrams_path = 'tempModel.h5'

# Generate the plot
plot_model(seq, to_file = model_diagrams_path + 'model_plot.png',
                show_shapes = True,
                show_layer_names = True)

# Show the plot here
img = mpimg.imread(model_diagrams_path + 'model_plot.png')
plt.figure(figsize=(30,15))
imgplot = plt.imshow(img)
```

```
framesArray                  =                np.load('/content/gdrive/My
Drive/dataFramesContours40x40.npy')
framesArrayShifted             =             np.load('/content/gdrive/My
Drive/dataFramesShiftedContours40x40.npy')


numSamples = len(framesArray)
print (numSamples)

plt.imshow(framesArray[0][0])
plt.gray()
plt.show()

plt.imshow(framesArrayShifted[0][0])
plt.gray()
plt.show()

movies  =  framesArray.reshape(framesArray.shape[0],  imgDepth,  imgRows,
imgCols, 1)
movies = movies.astype('float32')
movies /= 255
print(movies.shape)

numSamplesShifted = len(framesArrayShifted)
print (numSamplesShifted)

moviesShifted  =  framesArrayShifted.reshape(framesArrayShifted.shape[0],
imgDepth, imgRows, imgCols, 1)
moviesShifted = moviesShifted.astype('float32')
moviesShifted /= 255
print(moviesShifted.shape)

hist = seq.fit(movies[:45], moviesShifted[:45], batch_size=5,
        epochs=50, validation_split=0.05)


plt.plot(hist.history['accuracy'])
plt.plot(hist.history['val_accuracy'])
plt.title('model accuracy')
plt.ylabel('accuracy')
plt.xlabel('epoch')
plt.legend(['train', 'val'], loc='upper left')
plt.show()

# summarize history for loss
plt.plot(hist.history['loss'])
plt.plot(hist.history['val_loss'])
plt.title('model loss')
plt.ylabel('loss')
plt.xlabel('epoch')
plt.legend(['train', 'val'], loc='upper left')

plt.show()

preds = seq.evaluate(x = movies[45:48], y = movies[46:49])
print("================================")
print ("Loss = " + str(preds[0]))
print ("Test Accuracy = " + str(preds[1]))
```

```
%%timeit

which = 48
track = movies[which][:10, ::, ::, ::]

"""
for i in range(9):
    plt.imshow(track[i, ::, ::, 0])
    plt.gray()
    plt.show()
"""
for j in range(20):
    new_pos = seq.predict(track[np.newaxis, ::, ::, ::, ::])
    new = new_pos[::, -1, ::, ::, ::]
    track = np.concatenate((track, new), axis=0)

# And then compare the predictions
# to the ground truth
from google.colab import files
import statistics

from fsm_colab import *
fsm = GestureFSM()


track2 = movies[which][::, ::, ::, ::]
for i in range(30):
    fig = plt.figure(figsize=(10, 5))

    ax = fig.add_subplot(121)

    if i >= 10:
        ax.text(1,   3,   '{0}{1}'.format('Prediction   frame   ',   i),
fontsize=20, color='r')
    else:
        ax.text(1,   3,   '{0}{1}'.format('Initial   trajectory',   i),
fontsize=20, color='r')

    toplot = track[i, ::, ::, 0]

    plt.imshow(toplot)
    ax = fig.add_subplot(122)
    plt.text(1, 3, 'Ground truth', fontsize=20)



    toplot = track2[i, ::, ::, 0]
    if i >= 2:
        toplot = movies[which][i - 1, ::, ::, 0]

        if i >= 10:
            dist_euclidean = np.sqrt(sum((track[i, ::, ::, 0] - track2[i,
::, ::, 0])**2)) / track[i, ::, ::, 0].size
            dist_euclidean = np.average(dist_euclidean)
            plt.text(24,       3,       '{0:.5f}'.format(dist_euclidean),
fontsize=20, color='g', )
```

```
            dist_manhattan = np.sum(abs(track[i, ::, ::, 0] - track2[i,
::, ::, 0])) / track[i, ::, ::, 0].size
            dist_manhattan = np.average(dist_manhattan)
            plt.text(24,       6,       '{0:.5f}'.format(dist_manhattan),
fontsize=20, color='g', )

            dist_ncc = np.sum( (track[i, ::, ::, 0] - np.mean(track[i, ::,
::, 0])) * (track2[i, ::, ::, 0] - np.mean(track2[i, ::, ::, 0])) ) / \
              ((track[i, ::, ::, 0].size - 1) * np.std(track[i, ::, ::,
0]) * np.std(track2[i, ::, ::, 0]) )
            dist_ncc = np.average(dist_ncc)
            plt.text(24,   9,   '{0:.5f}'.format(dist_ncc),   fontsize=20,
color='g', )

            if dist_ncc >= 0.03:
                signal = True
                plt.text(24, 12, fsm.counter, fontsize=20, color='r', )
            else:
                signal = False
            fsm.send(signal)
            plt.text(24, 12, fsm.counter, fontsize=20, color='r', )

        if i == 10:
          np.save('track', track)
          files.download('track.npy')
          np.save('track2', track2)
          files.download('track2.npy')

    plt.imshow(toplot)
    plt.savefig('%i_animate.png' % (i + 1))
    #files.download('%i_animate.png' % (i + 1))
```