УДК 004.932.2

DOI: 10.20535/2523-4455.2019.24.6.197449

Удосконалений метод визначення положення суглобових з'єднань скелету людини на відеопослідовностях

Солдатов[f] Д. В., ORCID 0000-0002-2194-7717
Варфоломєєв[s] А. Ю., к.т.н., ORCID 0000-0002-6990-7140
Кафедра конструювання електронно-обчислювальної апаратури
Факультет електроніки
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Київ, Україна

***Анотація*—В роботі запропоновано ряд удосконалень методу визначення положення суглобових з'єднань скелету людини на відеопослідовностях з метою підвищення точності прогнозування положення людини у просторі. Це досягається за рахунок застосування наступних нововведень: врахування інформації про кути переміщення та наближення чи віддалення людини, що дозволяє розрізняти рухи, які схожі у відцентрованих кадрах, але відрізняються переміщенням; використання адаптивного розміру вікна для розрахунку HOG3D ознак; використання нейронної мережі для екстраполяції положень суглобових з'єднань у просторі у випадку відсутності або недостатньої точності прогнозування. Експериментальна перевірка, проведена на наборі даних HumanEva-1, показала підвищення в середньому на 11 пікселів точності локалізації суглобових з'єднань при застосуванні запропонованих модифікацій та підтвердила перспективність використання удосконаленого методу для подальшого вирішення задачі розпізнавання рухів.**

Ключові слова — розпізнавання рухів; прогнозування; CNN; HOG3D

I. Вступ

Надійне розпізнавання рухів людей має широке коло застосувань, включаючи ігри, взаємодію людини та комп'ютера, сферу безпеки та охорони здоров'я. В останні роки дослідниками комп'ютерної графіки та комп'ютерного зору розроблено нові алгоритми зйомки руху, які працюють на все більш простому апаратному забезпеченні та з набагато меншими обмеженнями, ніж раніше. Ці алгоритми не потребують спеціальних костюмів, щільних масивів камер, запису в студії або маркерів. Достатньо лише однієї відеокамери [1, 2, 3].

Вхідними даними задачі розпізнавання рухів є відео з людьми, що рухаються, вихідними – промарковані області з виявленими ключовими точками, які описують положення частин тіла та пояснення, який саме рух чи дія мають місце на відео (ходьба, біг, стрибки, різноманітні жести) [4, 5].

Розпізнавання рухів може виконуватись як апряму з RGB-зображення або відео з використанням різних варіацій згорткових нейронних мереж [6, 7], так і з попереднім представленням людини у вигляді скелету [8, 9].

Розпізнавання дій, що базується на скелетах, привертає все більше уваги, оскільки воно стійке до зміни масштабу тіла, швидкості руху, точок зору камери, фону та різноманітних перешкод. Воно також потребує менше ресурсів для обчислення, оскільки дані про скелети представляють людське тіло лише як послідовність координат ключових точок – основних суглобових з'єднань тіла (рис. 1) [10]. Цієї інформації зазвичай цілком достатньо для подальшого розпізнавання рухів [11, 12].

В роботі [11] авторами показано важливість поєднання інформації про суглоби та інформацію про кістки разом для розпізнавання дій, що базується на скелеті. Кістка представляється як різниця координат між двома з'єднаними суглобами. Тут, зокрема, дані 3D-скелету представляються наступним чином: суглоб представлений точкою з трьома координатами (x, y, z), тоді два суглобові з'єднання $j_1 = (x_1, y_1, z_1)$ і $j_2 = (x_2, y_2, z_2)$ утворюють кісткове з'єднання, що є вектором, який розраховується як покоординатна різниця двох точок суглобів, тобто: $b_{j1, j2} = (x_1 - x_2, y_1 - y_2, z_1 - z_2)$.

Для згаданих вище методів розпізнавання рухів на основі скелетного представлення важливим є точне визначення координат суглобових з'єднань. В той же час, існуючі для оцінювання координат рішення наразі не є ідеальними, а їх точність та надійність може бути підвищена, про що детальніше йтиметься в подальших розділах роботи.

DOI: 10.20535/2523-4455.2019.24.6.197449

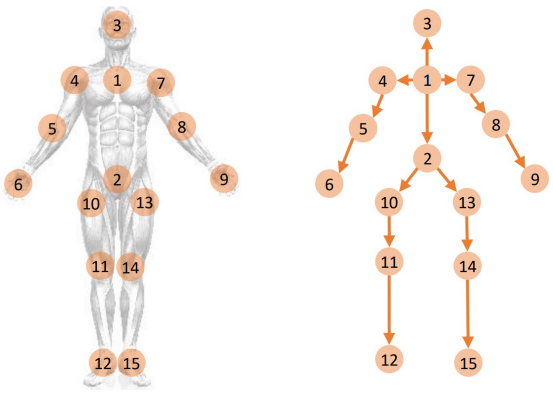

Рис. 1 Представлення людини у вигляді скелету (суглобів і кісток)

Мета даної роботи якраз і полягає у підвищенні точності прогнозування положення людини у просторі шляхом удосконалення та реалізації методу визначення положення суглобових точок скелету людини на відеопослідовностях, що може в подальшому використовуватись для реалізації систем розпізнавання рухів.

II. Аналіз існуючих методів

Підходи прогнозування положення людини можна розділити на дві основні категорії, залежно від того, базуються вони на нерухомих зображеннях чи на послідовностях зображень.

A. Прогнозування положення людини на одиночних зображеннях

Ранні підходи, як правило, покладалися на генеративні моделі пошуку стану простору для правдоподібної конфігурації скелета [13, 14]. Ці методи залишаються конкурентоспроможними за умови забезпечення достатньої якості ініціалізації. Наступні підходи [15, 16] розширюють двовимірну структуру [17] у тривимірну область. Однак, крім їх високої обчислювальної вартості, вони, як правило, мають труднощі з точною локалізацією рук людей, оскільки відповідні сигнали слабкі, і їх легко сплутати з фоном [18].

Навпаки, підходи, засновані на диференційній регресії [19, 20], будують пряму відповідність між зображенням та положенням скелету в просторі. Вони показали свою ефективність, особливо якщо є великий набір даних для навчання [21]. У цьому контексті функції кодування глибини [22] та інформації про частини тіла [23] виявились ефективними для підвищення точності прогнозування положення людини.

B. Прогнозування положення людини на послідовностях зображень

Такі підходи також розділяються на дві основні групи. Перша передбачає відстеження від кадру до кадру та динамічні моделі, на основі залежності Маркова між послідовними кадрами. Основний їх недолік полягає в тому, що вони потребують ініціалізації та не можуть відновитись після відмов.

Для усунення зазначених вище недоліків друга група зосереджується на виявленні кандидатів в окремих кадрах з подальшим зв'язком у часовій послідовності. Наприклад, у [24] початкові оцінки позиції уточнюються за допомогою двовимірних оцінок на основі коротких послідовностей зображень. В роботі [25] інтегрується відновлення положення тіла на кожному кадрі з K найкращими траєкторіями та адаптацією текстури моделі. В роботі [26] використовується щільний оптичний потік для виявлення зв'язку моделей положення у сусідніх кадрах. На відміну від цих підходів у [27] часова інформація фіксується раніше: просторово-часові ознаки вилучаються з коротких послідовностей і за допомогою регресійних моделей перетворюються у тривимірні пози. В якості просторово-часових ознак використовується тривимірний дескриптор HOG, описаний в [28].

Сильною стороною 3D-дескриптора HOG є те, що він відносно просто дозволяє кодувати інформацію як про зовнішній вигляд, так і про рух. На відміну від двовимірного варіанта дескриптора HOG, він не тільки зберігає інформацію про зовнішній вигляд, але й кодує інформацію про рух, обчислюючи часові градієнти. Альтернативою такому кодуванню інформації про рух було б чітке відстеження частин тіла в просторово-часовому об'ємі.

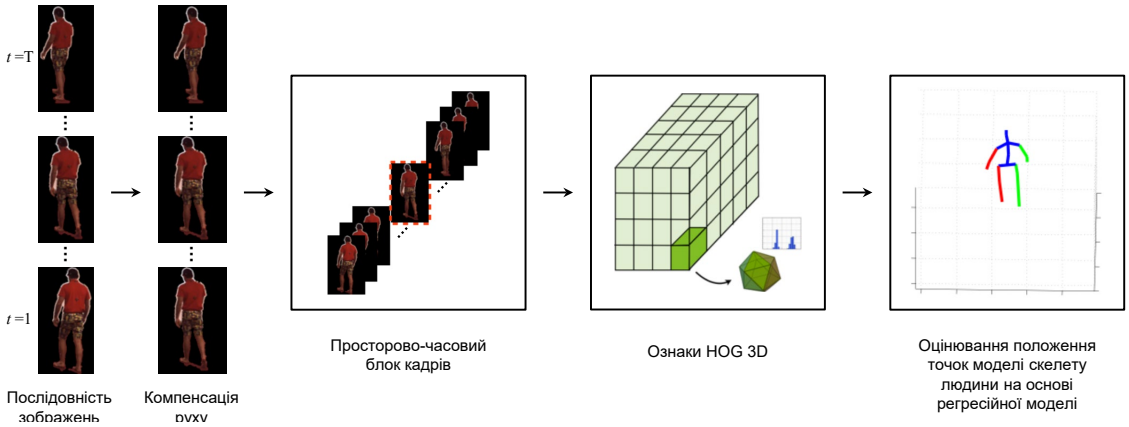

Рис. 2 Реалізація методу визначення положення суглобових з'єднань скелету людини на відеопослідовностях

DOI: 10.20535/2523-4455.2019.24.6.197449

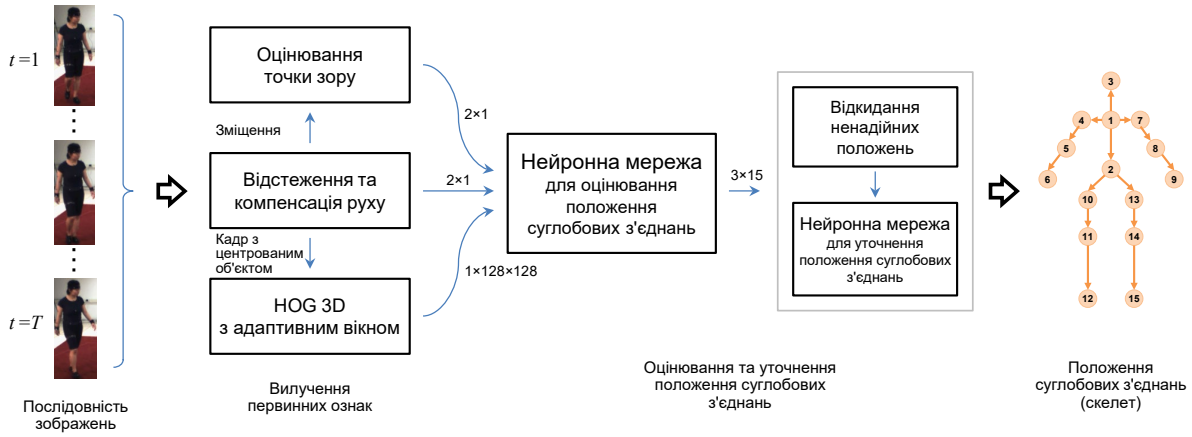

Рис. 3 Реалізація запропонованого методу визначення положення суглобових з'єднань скелету людини на відеопослідовностях

Положення тіла представляється у вигляді скелету, схожого на той, що показано на рис. 1. Розташування суглобових з'єднань у такому представленні визначається відносно деякого кореневого вузла. Таке представлення добре підходить для побудови регресійних моделей і не вимагає знань про точні пропорції тіла. Хоча воно і не є орієнтаційно-інваріантним, додаткове використання часової інформації нівелює цей недолік. Реалізацію зазначеного підходу наведено на рис. 2.

На вхід подається послідовність зображень, для якої виконується компенсація руху. Це необхідно, так як для згаданих вище дескрипторів HOG3D, часові ознаки повинні відповідати конкретним частинам тіла. Це означає, що особа повинна залишатися центрованою від кадру до кадру. Без цього градієнти виявляються «розпорошеними», що знижує якість дескриптора [27].

За допомогою алгоритму відстеження будуються прямокутники навколо об'єкту. Потім зображення зсуваються так, що об'єкт залишається в центрі. Далі вилучається піраміда ознак тривимірного HOG з фіксованим розміром часового вікна (тобто фіксованою кількістю кадрів). Остання процедура визначає положення скелету у просторі, використовуючи регресійну модель [27].

Розглянутий метод володіє одними з найкращих у своєму класі характеристиками точності та надійності виявлення суглобових з'єднань. Водночас, його показники можуть бути підвищені ще за рахунок запропонованих в наступному розділі вдосконалень.

III. Сутність запропонованого методу

Запропонований в цій статті метод (рис. 3) базується на підході з роботи [27]. Вхідна послідовність зображень передається у процедуру відстеження і компенсації руху, в якій за допомогою алгоритму відслідковування визначаються рамки, що оточують об'єкт, та відбувається його центрування в області кадру. Дані про рух також передаються до процедури оцінювання точки зору та до нейронної мережі, а кадри з відцентрованим об'єктом до блоку HOG3D. Для відслідковування і компенсації руху використовується вдосконалений підхід на базі оптичного потоку, що описаний в роботі [29].

Процедура оцінювання точки зору, використовуючи дані про компенсацію руху, розраховує кути напрямку переміщення людини в кадрі та слідкує за зміною її розміру, визначаючи тим самим наближається чи віддаляється людина від камери. Далі ці дані передаються до нейронної мережі, дозволяючи їй точніше оцінювати координати суглобових з'єднань скелету.

Процедура обчислення ознак HOG3D вилучає піраміду ознак тривимірного HOG і передає її до нейронної мережі. В цій роботі ознаки HOG3D розраховуються у часовому вікні адаптивної тривалості, яка залежить від швидкості руху людини у кадрі. Такий підхід, як показано в роботі [20], дозволяє підвищити точність прогнозування положення людини.

Врахування інформації про кути переміщення, наближення та віддалення людини дозволяє краще розрізняти схожі рухи на кадрах, в яких положення людини відцентровано, проте сам цей рух відрізняється.

Нейронна мережа, що виконує пошук положень суглобових з'єднань, навчена за даними, які розраховуються блоками оцінювання точки зору, відстеження і компенсації руху, а також блоку обчислення HOG3D ознак. Вказана мережа є згортковою та має стандартну, схожу на описану в роботі [30] структуру. Вона, зокрема, складається з трьох згорткових шарів (convolutional layers), за кожним з яких слідує шар об'єднання (pooling layer). Останній об'єднуючий шар підключається до каскаду з трьох повно зв'язних шарів. За винятком останнього шару, де використовується лінійна активація, кожен шар повнозв'язної частини використовує функцію активації ReLU. Ознаки HOG3D поступають до згорткових шарів, а дані від процедур відстеження та оцінювання точки зору одразу передаються на повнозв'язні шари мережі. Структура мережі наведена на рис. 4. При навчанні використовувався класичний алгоритм зворотного розповсюдження помилки зі стохастичним градієнтним спуском. Для того, щоб уникнути перенавчання використовувалась техніка відкидання

DOI: 10.20535/2523-4455.2019.24.6.197449

(dropout) [31], з коефіцієнтом відкидання 0.25. Навчання проводилось на наборі даних HumanEva-1 [5].

Таблиця 1 Результати оцінювання похибок положення об'єкту на тестовому наборі HumanEva-1

Реалізація	Біг	Ходьба	Середнє
Базова	59,4	52,1	55,75
Модифікована	46,2	41,6	43,9

Процедура уточнення положень суглобових з'єднань виконує дві функції. По-перше, нею визначаються помилкові положення, якими визнаються ті, в яких відбулась або занадто різка зміна координат, або для яких відсутні надійні прогнози нейронної мережі. По-друге, якщо положення деякого суглобового з'єднання було визнано знайденим ненадійно, то для такого з'єднання застосовується екстраполяція координат на основі раніше отриманих для нього положень, для цього застосовується додаткова нейронна мережа. Дана мережа є досить простою та складається з чотирьох повнозв'язних шарів. Її навчання також виконується на навчальній множині HumanEva-1. Таким чином, на виході маємо неперервну послідовність прогнозованих положень суглобів у просторі, що відповідно підвищує стабільність та зменшує помилку оцінювання координат.

Перевірка методу виконувалась на тестовій частині набору даних HumanEva-1 для ходьби та бігу. Результати перевірки зведені до табл. 1. Для порівняння у першому рядку табл. 1 також наведено результати роботи базової реалізації методу, що не використовує блок оцінювання точки зору та додаткову нейронну мережу для екстраполяції. При цьому в якості міри оцінювання точності роботи систем використовується середня евклідова відстань між прогнозованим системою та реальними еталонними положеннями суглобових з'єднань:

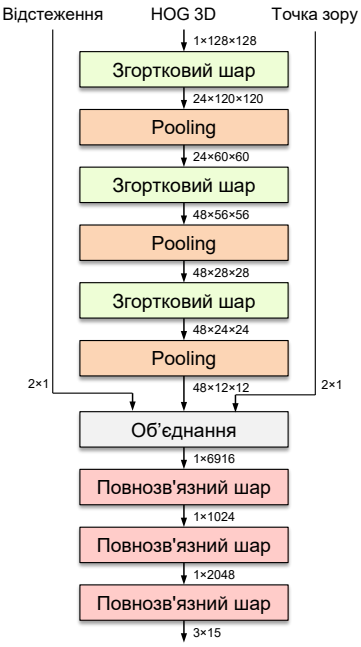

Рис. 4 Архітектура нейронної мережі для первинного прогнозування положення суглобів

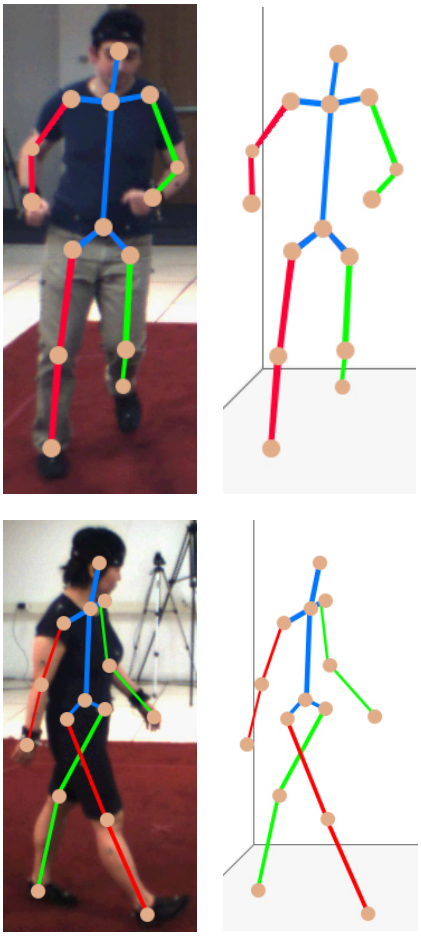

Рис. 5 Результат роботи запропонованого методу, для бігу та ходьби

$$E = \frac{1}{M \cdot T} \sum_{t=1}^{T} \sum_{i=1}^{M} \| j_{i,t} - \hat{j}_{i,t} \|,$$

де $j_{i,t}$, $\hat{j}_{i,t}$ – координати i-го суглобу знайденого системою та еталонні координати цього ж суглобу у наборі даних відповідно на t-му кадрі відеопослідовності; $\| \cdot \|$ позначає евклідову норму; M – кількість суглобових з'єднань ($M = 15$); T – кількість кадрів у відеопослідовності або відеопослідовностях, на яких здійснюється оцінювання.

Із наведених у табл. 1 результатів видно, що запропоновані вдосконалення дозволяють зменшити похибку оцінювання координат приблизно на 11 пікселів, що підтверджує доцільність запропонованих модифікацій. На рис. 5 наведено приклад результату роботи методу.

Висновки

В роботі запропоновано модифікований метод визначення положення суглобових з'єднань скелету людини на відеопослідовностях, що відрізняється від відомих рішень використанням додаткових складових: процедури оцінювання точки огляду, яка враховує кути напрямку переміщення об'єкта у кадрі та слідкує за його розміром, а також нейронної мережі для екстраполяції помилкових оцінок положень

DOI: 10.20535/2523-4455.2019.24.6.197449

суглобових з'єднань. Під час експериментальної перевірки на тестовому наборі даних HumanEva-1 встановлено, що вдосконалений метод забезпечує на 11 пікселів меншу похибку оцінювання координат у порівнянні з реалізацією, яка не використовує запропоновані модифікації.

Основу майбутніх досліджень можуть скласти розширення експериментальної перевірки іншими наборами даних та використання запропонованого методу для розроблення системи розпізнавання рухів.

Перелік посилань

[1] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-toend Recovery of Human Shape and Pose. CVPR, [Online]. **Available**: https://arxiv.org/abs/1712.06584, 2018. [Accessed 29 02 2020]

[2] D. Xiang, H. Joo, and Y. Sheikh. Monocular Total Capture: Posing Face, Body, and Hands in the Wild. [Online]. **Available**: https://arxiv.org/abs/1812.01598, 2018. [Accessed 29 02 2020]

[3] D.Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.P. Seidel, H. Rhodin, G. Pons-Moll, C. Theobalt, XNect: Real-time Multi-person 3D Human Pose Estimation with a Single RGB Camera, [Online] **Available**: https://arxiv.org/abs/1907.00837 2019 [Accessed 29 02 2020]

[4] C. Ionescu, D. Papava, V. Olaru and C. Sminchisescu, Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, No. 7, July 2014.

[5] L. Sigal, A. Balan and M. J. Black, "HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion," *International Journal of Computer Vision (IJCV)*, vol. 87, pp. 4–27, 2010. **DOI:** 10.1007/s11263-009-0273-6

[6] S. Li and A. B. Chan, "3D Human Pose Estimation from Monocular Images with Deep Convolutional Network," *Asian Conference on Computer Vision (ACCV)*, 2014. **DOI:** 10.1007/978-3-319-16808-1_23

[7] N.C. Camgoz, S. Hadfield, O. Koller and R. Bowden, "Using convolutional 3D neural networks for userindependent continuous gesture recognition," *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 49–54, 2016. **DOI:** 10.1109/ICPR.2016.7899606

[8] C. Cao, C. Lan, Y. Zhang, W. Zeng, H. Lu and Y. Zhang. "Skeleton-Based Action Recognition with Gated Convolutional Neural Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 11, pp. 3247–3257, 2018. **DOI:** 10.1109/TCSVT.2018.2879913

[9] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 579–583, 2015. **DOI:** 10.1109/ACPR.2015.7486569

[10] Skeleton-Based Action Recognition with Directed Graph Neural Networks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) 2019.* **DOI**: 10.1109/CVPR.2019.00810

[11] L. Li, W. Zheng, Z. Zhang, Y. Huang, and L. Wang, "Skeleton-Based Relational Modeling for Action Recognition" [Online]. **Available**: https://arxiv.org/abs/1805.02556, 2018. [Accessed 29 02 2020]

[12] L. Shi, Y. Zhang, J. Cheng, and H. Lu. NonLocal Graph Convolutional Networks for Skeleton-Based Action Recognition. [Online] **Available**: https://arxiv.org/abs/1805.07694, May 2018. [Accessed 29 02 2020]

[13] R. Urtasun, D. Fleet, and P. Fua, "3D People Tracking with Gaussian Process Dynamical Models," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006. **DOI:** 10.1109/CVPR.2006.15

[14] C. Sminchisescu and B. Triggs, "Covariance Scaled Sampling for Monocular 3D Body Tracking," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001. **DOI:** 10.1109/CVPR.2001.990509

[15] M. Burenius, J. Sullivan and S. Carlsson, "3D Pictorial Structures for Multiple View Articulated Pose Estimation," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. **DOI:** 10.1109/CVPR.2013.464

[16] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab and S. Ilic, "3D Pictorial Structures for Multiple Human Pose Estimation" *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. **DOI:** 10.1109/CVPR.2014.216

[17] P. Felzenszwalb, R. Girshick, D. McAllester and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 32, no. 9, pp. 1627–1645, 2010. **DOI:** 10.1109/TPAMI.2009.167

[18] B. Sapp, A. Toshev and B. Taskar, "Cascaded Models for Articulated Pose Estimation," *Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science*, vol. 6312, pp. 406–420, 2010. **DOI:** 10.1007/978-3-642-15552-9_30

[19] A. Agarwal and B. Triggs, "3D Human Pose from Silhouettes by Relevance Vector Regression," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004. **DOI:** 10.1109/CVPR.2004.1315258

[20] L. Sigal, A. Balan and M. J. Black, "Combined Discriminative and Generative Articulated Pose and Non-rigid Shape Estimation," *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[21] C. Ionescu, I. Papava, V. Olaru and C. Sminchisescu. "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 36, no. 7, pp. 1325–1339, 2014. **DOI:** 10.1109/TPAMI.2013.248

[22] J. Shotton, A. Fitzgibbon, M. Cook and A. Blake, "Real-Time Human Pose Recognition in Parts from a Single Depth Image," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. **DOI:** 10.1109/CVPR.2011.5995316

[23] C. Ionescu, J. Carreira and C. Sminchisescu, "Iterated Second-Order Label Sensitive Pooling for 3D Human Pose Estimation," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. **DOI:** 10.1109/CVPR.2014.215

[24] M. Andriluka, S. Roth and B. Schiele, "Monocular 3D Pose Estimation and Tracking by Detection," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. **DOI:** 10.1109/CVPR.2010.5540156

[25] M. Hofmann and D. M. Gavrila, "Multi-view 3D Human Pose Estimation in Complex Environment," *International Journal of Computer Vision (IJCV)*, vol. 96, pp. 103–124, 2012. **DOI:** 10.1007/s11263-011-0451-1

[26] S. Zuffi, J. Romero, C. Schmid and M. J. Black, "Estimating Human Pose with Flowing Puppets," *IEEE International Conference on Computer Vision (ICCV)*, 2013. **DOI:** 10.1109/ICCV.2013.411

[27] B. Tekin, X. Sun, X. Wang, V. Lepetit and P. Fua, "Predicting People's 3D Poses from Short Sequences" [Online]. **Available**: https://arxiv.org/abs/1504.08200, 2018. [Accessed 29 02 2020]

[28] D. Weinland, M. Ozuysal and P. Fua, "Making Action Recognition Robust to Occlusions and Viewpoint Changes," *Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science*, vol. 6313, pp. 635–648, 2010. **DOI:** 10.1007/978-3-642-15558-1_46

[29] D. Park, C. L. Zitnick, D. Ramanan and P. Dollar, "Exploring Weak Stabilization for Motion Feature Extraction," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. **DOI:** 10.1109/CVPR.2013.371

[30] S. Li and A.B. Chan. 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network. In ACCV, 2014 **DOI**: 10.1007/978-3-319-16808-1_23

Надійшла до редакції 02 грудня 2019 р.

[31] Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Improving neural networks by preventing co-adaptation of feature detectors. CoRR (2012) [Online]. **Available:** https://arxiv.org/abs/1207.0580 [Accessed 29 02 2020]

DOI: 10.20535/2523-4455.2019.24.6.197449

УДК 004.932.2

Усовершенствованный метод определения положения суставных соединений скелета человека на видеопоследовательностях

Солдатов[f] Д. В., ORCID 0000-0002-2194-7717
Варфоломеев[s] А. Ю., к.т.н., ORCID 0000-0002-6990-7140
Кафедра конструирования электронно-вычислительной аппаратуры
Факультет электроники
Национальный технический университет Украины
«Киевский политехнический институт имени Игоря Сикорского»
Киев, Украина

***Аннотация*—В работе предложено ряд усовершенствований метода определения положения суставных соединений скелета человека на видеопоследовательностях с целью повышения точности прогнозирования положения человека в пространстве. Это достигается за счет применения следующих нововведений: учета информации об углах перемещения и приближения или отдаления человека, что позволяет распознавать движения, в отцентрированных кадрах, но отличающихся перемещением; использованием адаптивного размера окна для расчета HOG3D признаков; использования нейронной сети для экстраполяции положений суставных соединений в пространстве при отсутствии или недостаточной точности прогнозирования. Экспериментальная проверка, проведенная на наборе данных HumanEva-1, показала увеличение точности локализации суставных соединений в среднем на 11 пикселей в случае применения предложенных модификаций и подтвердила перспективность использования усовершенствованного метода для дальнейшего решения задачи распознавания движений.**

Ключевые слова — распознавание движений; прогнозирование; CNN; HOG3D

UDC 004.932.2

DOI: 10.20535/2523-4455.2019.24.6.197449

The Modified Method for Position Estimation of Human Body Joints in Video Sequences

D. V. Soldatov[f], ORCID 0000-0002-2194-7717
A. Y. Varfolomieiev[s], PhD, ORCID 0000-0002-6990-7140
Department of Design of Electronic Computing Equipment
Faculty of Electronics
National technical university of Ukraine "Igor Sikorsky Kyiv polytechnic institute"
Kyiv, Ukraine

Abstract*—Reliable recognition of human movements has a wide range of applications, including games, human-computer interaction, security and healthcare. In recent years, computer graphics and computer vision researchers have developed plenty of new motion-capture algorithms that operate on simpler hardware and with far fewer limitations than before. The objective of this paper is to improve the accuracy of estimation of human skeleton joints positions in video sequences. Particularly the proposed method in this paper consists of five blocks. The input sequence of images is fed to the tracking and motion compensation unit where the tracking algorithm determines the object displacement and centers it within the frame. The motion information is also propagated to the additional unit of point-of-view estimation. This unit calculates the motion angles in the frame and monitors the object size, thus determining whether the object is approaching or moving away from the camera, and then feeds these data to the neural network. The network consists of three convolutional layers. Each convolutional layer is followed by a pooling layer. The last pooling layer connects to the cascade of three fully connected layers. All activation functions in these layers are the ReLU ones, except the last layer, where the linear activation is used. The HOG3D features treated as the input of the first convolutional layer. The data from the point-of-view, tracking and motion compensation unit goes directly to the input of fully connected layers. To cope with inaccurate or undetected joints positions, the method uses the additional procedure, which determines unreliable joints and extrapolates their new positions from the previous ones using the additional neural network. It is assumed that this structure of the method improves the position prediction accuracy due to the following reasons: taking into account the information about motion angles and zooming allows to distinguish movements that are similar in centered frames but different in displacement; using of adaptive window size for HOG3D features; using the neural network to extrapolate the positions of joints in case of absence of the prediction or in case of its low accuracy. Experiments on the HumanEva-1 dataset confirmed that the suggested modifications permit achieving higher accuracies, and thus the prospect of the use of proposed modified method to predict the body position in motion recognition systems.

Keywords — Motion recognition; pose prediction; CNN; HOG3D.