ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2023, том 514, № 2, с. 364–374

УДК 004.93

ИССЛЕДОВАНИЕ НЕЙРОСЕТЕВЫХ АЛГОРИТМОВ ПРОГНОЗИРОВАНИЯ ДВИЖЕНИЙ ЧЕЛОВЕКА НА БАЗЕ LSTM И ТРАНСФОРМЕРОВ

© 2023 г. С. В. Жиганов[1,*], Ю. С. Иванов[1,], Д. М. Грабарь[1,***]**

Представлено академиком РАН А.И. Аветисяном
Поступило 02.09.2023 г.
После доработки 15.09.2023 г.
Принято к публикации 24.10.2023 г.

Решена задача прогнозирования положения человека на будущих кадрах видеопотока и проведены глубокие экспериментальные исследования по применению традиционных и SOTA блоков для данной задачи. Представлены оригинальная архитектура KeyFNet и ее модификации, основанная на трансформеных блоках, способная предсказать координаты в видеопотоке на 30, 60, 90 и 120 кадров вперед с высокой точностью. Новизна состоит в применении комбинированного алгоритма на основе нескольких блоков FNet с быстрым преобразованием Фурье в качестве механизма внимания, конкатенирующих координаты ключевых точек. Проведенные эксперименты на Human3.6M и на собственных реальных данных подтвердили эффективность предложенного подхода на базе блоков FNet, в сравнении с традиционным подходом, основанным на LSTM. Предлагаемый алгоритм соответствует по точности передовым моделям, но превосходит их по скорости и использует меньше ресурсов для вычисления и может быть применен в коллаборативных робототехнических решениях.

Ключевые слова: прогнозирование ключевых точек, трансформеры, коллаборативные робототехнические системы, глубокое обучение

DOI: 10.31857/S2686954323601689, **EDN:** CXAFDM

1. ВВЕДЕНИЕ

В современных робототехнических системах, особенно в контексте коллаборативных операций, взаимодействие между человеком-оператором и роботом играет решающую роль в обеспечении эффективности и безопасности задач, которые выполняются совместно. Для совместной работы человека и робота эффективное взаимодействие требует точного прогнозирования движений и положения оператора для обеспечения плавной и безопасной кооперации. Это позволяет роботам адекватно воспринимать инструкции и действия оператора, а также адаптироваться к изменениям в окружающей среде.

Стоит учитывать, что задача обнаружения и предсказания поведения оператора коллаборативной робототехнической системы особенно сложна из-за неоднородности объектов, имеющих различные и потенциально сложные формы, а также трудностей, возникающих из-за фоновых помех и частичных перекрытий между объектами (окклюзий).

Оценка будущего положения и позы человека определяется как предсказание местоположения ключевых точек человека на будущих кадрах, с учетом наблюдаемых (обнаруженных) точек прошлых кадров.

При этом задача осложняется необходимостью учитывать взаимосвязь сложных пространственно-временных взаимодействий между частями тела (например, руками, ногами, позвоночником).

Существуют подходы, основанные на прогнозировании траектории движения участков изображения [6, 6], однако их общим недостатком является невозможность учета особенностей движений человека.

В работе [6] авторы используют единственное статичное изображение для предсказания последовательности будущих расположений ключевых точек на базе глубокой нейронной сети со сверточными слоями. Отличительной особенностью подхода является наличие дополнительного ком-

[1]*Комсомольский-на-Амуре государственный университет, Комсомольск-на-Амуре, Россия*
**E-mail: id_zero@list.ru*
***E-mail: ivanov_ys@icloud.com*
****E-mail: gorbat308@yandex.ru*

понента для дальнейшего преобразования каждой прогнозируемой позы из 2D-пространства в 3D-пространство.

В последние годы наблюдается значительный прогресс в области использования традиционных глубоких нейросетевых архитектур [6, 6], и таких как трансформеры, в задачах компьютерного зрения и обработки данных [6].

Традиционно для задач обработки временных рядов применяются архитектуры на базе блоков LSTM или RNN-подобных блоков [6]. Авторы [6] решают задачу не только прогнозирования будущего положения человека, но и генерации видео с реалистичными движениями.

Трансформеры показали выдающиеся результаты в области обработки последовательностей, что стимулирует их применение и в задачах, связанных с анализом движений, и позиции человека. При этом в области компьютерного зрения трансформеры показываю SOTA результаты или близкие к ним. Так, например, авторы предложили архитектуру Swin Transformer [6], которая превзошла применяемый ранее Vision Transformer.

Целью данной статьи являются анализ различных подходов к задаче прогнозирования ключевых точек человека, оценка их применимости в коллаборативных сценариях и предложение нового модифицированного метода, основанного на трансформерах.

В следующих разделах статьи приведены постановка задачи прогнозирования положения человека с использованием ключевых точек, описание используемых метрик и наборов данных, а также представлен наш подход к решению поставленной задачи.

Описана подробная архитектура предлагаемого метода, а также ее модификации с использованием трансформеров. Приведены результаты экспериментов, их анализ и перспективы дальнейших исследований в данной области. Мы также рассмотрим вопросы влияния особенностей биомеханики человеческого тела и скорости перемещения отдельных точек на точность прогнозирования предлагаемых алгоритмов.

Данная статья может представлять весомый вклад в развитие коллаборативных робототехнических систем, обеспечивая более точное и надежное взаимодействие между человеком-оператором и роботами за счет использования передовых SOTA методов анализа данных, основанных на трансформерах.

2. ПОСТАНОВКА ЗАДАЧИ

Пусть имеется фиксированный набор кадров непрерывного видеопотока наблюдаемой сцены, в которой могут присутствовать люди $\mathbf{V} = (\mathbf{I}_{t-N} \ldots \mathbf{I}_t)$, где t – текущий момент времени. Размер фиксированного набора кадра N назовем окном.

Присутствие человека и его положение описываются вектором, содержащим координаты ключевых точек в пространстве наблюдаемой сцены $\mathbf{K} = (x_1, y_1, x_2, y_2, \ldots, x_n, y_n)$ где n – количество ключевых точек человека в соответствии с выбранным стандартом.

Таким образом, наблюдаемая сцена может быть представлена в виде временного ряда $\mathbf{K}_{hist} = (\mathbf{K}_{t-N}, \mathbf{K}_{t-N+1}, \mathbf{K}_{t-N+2}, \ldots, \mathbf{K}_t)$ фиксированного размера N, начинающегося с исторических данных и заканчивающегося текущим моментом t.

Требуется по имеющемся историческим данным $\mathbf{K}_{hist}$ предсказать траекторию движения человека на M кадров вперед $\mathbf{K}_{forecast} = (\mathbf{K}_{t+1}, \ldots, \mathbf{K}_{t+M})$, в соответствии с заданным критерием $P(\mathbf{K}_{pred})$ минимизирующим вероятность ошибки, где $\mathbf{K}_{pred}$ – вектор координат ключевых точек каждого предсказанного кадра.

Таким образом, необходимо найти отображение $\mathbf{F} : \mathbf{K}_{hist} \to \mathbf{K}_{forecast}$, при котором F является набором функций и алгоритмов, на которые накладываются ограничения по ресурсоемкости и быстроте для работы в реальном режиме времени.

Под критериями качества будем понимать следующие, общепринятые для данной задачи, методы оценки эффективности моделей машинного обучения на основе регрессии:

• среднеквадратическая ошибка прогноза *MSE* (Mean Squared Error) – это среднее значение квадрата разницы между фактическими координатами точки и значениями, предсказанными алгоритмом:

$$MSE = \frac{\sum_{i=1}^{n} (y_i^{\text{true}} - y_i^{\text{pred}})}{n}, \quad (1)$$

где n – количество примеров в обучающей выборке, y_i^{true} – истинное значение, y_i^{pred} – прогнозируемое значение.

• среднеквадратичная ошибка (Root Mean Square Error) – это значение ошибки, полученное вычислением квадратного корня из *MSE*:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i^{\text{true}} - y_i^{\text{pred}})}{n}}, \quad (2)$$

где n – количество примеров в обучающей выборке, y_i^{true} – истинное значение, y_i^{pred} – прогнозируемое значение.

• коэффициент детерминации (R^2) – это коэффициент, который показывает, насколько хоро-

шо модель соответствует зависимым переменным. R^2 можно интерпретировать как измерение количества отклонений в прогнозах, объясненных набором данных. Результат может принимать значения от 0 до 1:

$$R^2 = 1 - \frac{\sum_i (y_i^{\text{true}} - y_i^{\text{pred}})^2}{\sum_i \left(y_i^{\text{true}} - \frac{\sum_i y_i^{\text{true}}}{n} \right)^2}, \quad (3)$$

где n – количество примеров в обучающей выборке, y_i^{true} – истинное значение, y_i^{pred} – прогнозируемое значение.

• средняя абсолютная ошибка MAE (Mean Absolute Error) – измеряет среднюю абсолютную величину между фактическими значениями и значениями, предсказанными регрессионной моделью:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y_i^{\text{true}} - y_i^{\text{pred}} \right|, \quad (4)$$

где n – количество примеров в обучающей выборке, y_i^{true} – истинное значение, y_i^{pred} – прогнозируемое значение.

При вычислении общей эффективности результат каждой из метрик усредняется по всем точкам.

Решение задачи прогнозирования положения человека на кадрах видеопотока разбивается на решение ряда подзадач:

1. Выполняется предобработка данных, включающая в себя нормализацию параметров.

2. Выполняется прогнозирование координат ключевых точек на заданный временной интервал с использованием глубоких нейронных сетей и трансформеров.

3. Выполняется переход к реальным координатам путем денормализации прогнозного вектора.

4. Усреднение прогнозных значений с использованием дискретных временных шагов для уменьшения ошибочно.

Общий алгоритм представлен на рис. 1.

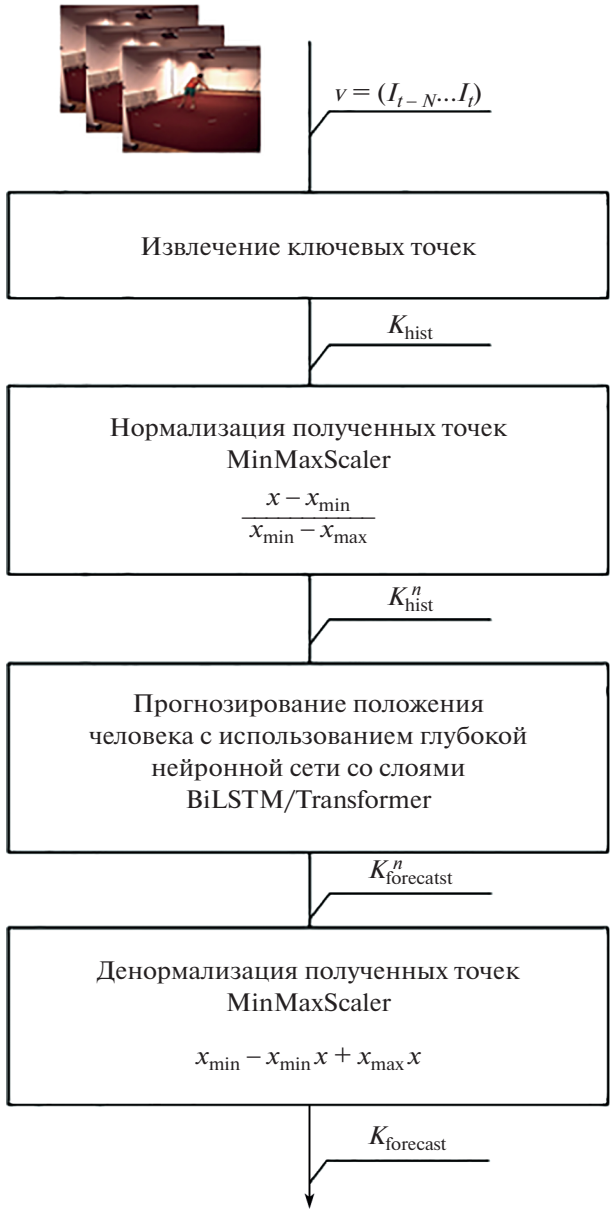

Рис. 1. Решение задачи прогнозирования положения человека на кадрах видеопотока.

Предобработка данных и нормализация признаков

Для обучения и тестирования использовался набор данных Human3.6M [6]. В наборе данных содержится 3.6 миллионов 3-х мерных позиций человека в 17 различных сценариях, записанных с привлечением 11 профессиональных актеров, из которых 6 женщин 5 мужчин. При формировании набора данных использовались 4 откалиброванные камеры с высоким разрешением, высокоскоростная система захвата движений. Каждый кадр имеет информацию о положении человека, пример аннотации и разметки кадра сцены представлены рис. 2.

В дальнейшем для локализации позиции человека на кадрах использовался алгоритм, представленный в статье [11].

Из набора данных сформировано 4 обучающих множества $\mathbf{D}^s = (\mathbf{K}_{hist}, \mathbf{K}_{t+s})$, где s – целевое положение человека на 30, 60, 90, 120 кадрах в будущем. Каждое обучающее множество $\mathbf{D}^s$ по сценам с действиями людей разделено на обучающую часть $\mathbf{D}^s_{\text{train}}$ 80% и тестирующую $\mathbf{D}^s_{\text{valid}}$ 20%.

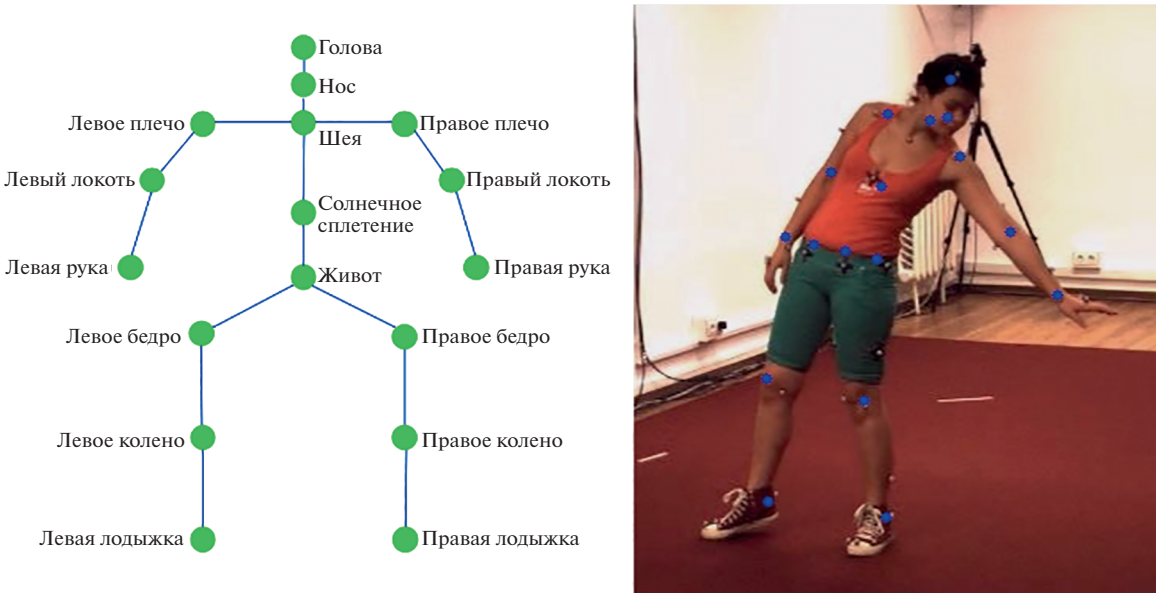

Рис. 2. Примеры аннотации и разметки кадров в наборе данных.

Значения параметров обучающего вектора имеют очень большой разброс (распределение) и масштаб, вследствие чего алгоритм машинного обучения может “предположить”, что один из признаков важнее другого, только исходя из их значений. Для устранения данного эффекта необходимо выполнить нормализацию. Нами предлагается:

(1) снижение масштаба пространства координат путем деления на коэффициент l.

(2) использование функции максимального и минимального масштабирования (MinMaxScaler), которая приводит значение каждого признака к диапазону от 0 до 1, предотвращая неправильное поведение классификатора.

Нормализация выполняется по следующей формуле:

$$\frac{x - x_{\min}}{x_{\min} - x_{\max}}, \quad (5)$$

где x – где значение признака, $x_{\min}$ – минимальное значение признака, $x_{\max}$ – максимальное значение признака.

Для восстановления признаков после выполнения алгоритма машинного обучения выполняется обратное преобразование:

$$x_{\min} - x_{\min}x + x_{\max}x, \quad (6)$$

где x – где значение признака, $x_{\min}$ – минимальное значение признака, $x_{\max}$ – максимальное значение признака.

3. ПРОГНОЗИРОВАНИЕ ПОЛОЖЕНИЯ ПОЗЫ ЧЕЛОВЕКА С ИСПОЛЬЗОВАНИЕМ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

Итоговая архитектура нейронной сети для оценки и прогнозирования положения человека на последующих кадрах представлена на рис. 3.

В процессе исследования нами был предложен ряд модификаций, повышающих качество прогнозирования. В табл. 1 представлены сводные результаты всех экспериментов для каждой точки прогнозирования: 30, 60, 90 и 120 кадров.

С левой стороны по вертикали приведены архитектуры и их модификации, а по горизонтали приведены окна прогнозирования и метрики.

В силу ограниченности пространства, полные данные по экспериментам приведены по ссылке [12].

Вариации модификаций блоков отмечены на рис. 4 соответствующими аннотациями. Опишем их подробнее.

Предлагаемая архитектура состоит из полносвязанных нейросетевых блоков $h_{\text{fcn}}^{\text{relu}}$ с функцией активации *Relu* и $h_{\text{fcn}}^{\text{linear}}$ с линейной функцией активации *linear* и блоков для выделения взаимосвязи во временных данных и координат h_{ts}

Базовая архитектура

Мы провели ряд экспериментов по выбору нейросетевой архитектуры для блоков h_{ts} и остановились на следующих:

• блок долгой краткосрочной памяти (LSTM), представляющий собой разновидность архитектуры рекуррентных нейронных сетей. LSTM показали хорошие результаты при решении задач, в

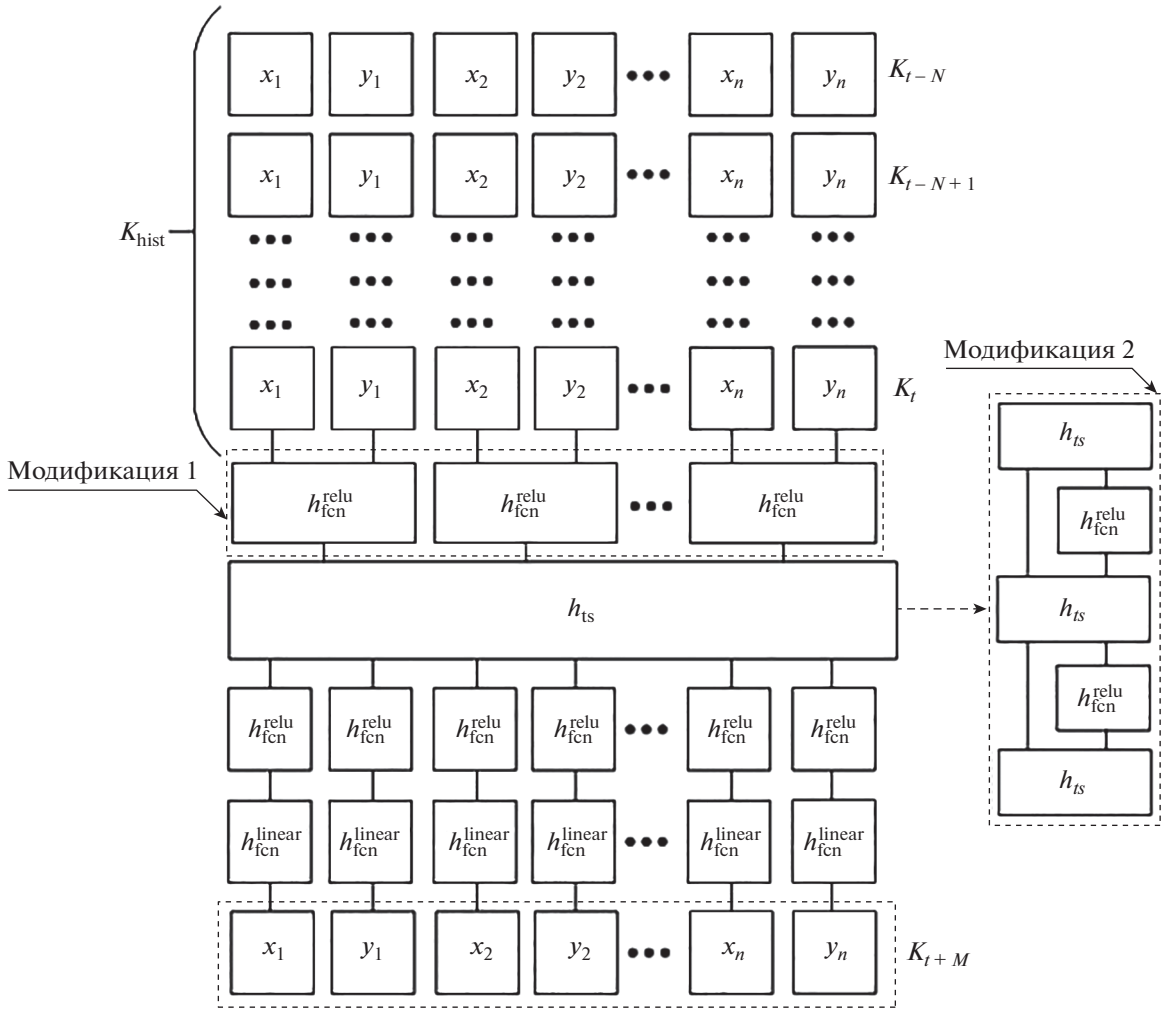

Рис. 3. Модифицированная архитектура для оценки и прогнозирования положения человека на будущих кадрах с блоками FNet/LSTM.

которых важные события разделены временными лагами с неопределенной продолжительностью и границами, что, в целом, соответствует задаче, решаемой в данной работе. Двунаправленная LSTM (Bi-LSTM) представляет собой модификацию классической сети, в которой входные данные передаются в обоих направлениях и могут использовать информацию с обеих сторон. Архитектурно Bi-LSTM, как правило, представляют двумя блоками LSTM для входа и выхода.

• блок FNet – разновидность архитектуры трансформера, [6], в которой используются быстрое преобразование Фурье и линейные преобразования в качестве блоков внимания. FNet представляет собой нормализованную по уровням архитектуру ResNet с несколькими уровнями, каждый из которых состоит из подуровня смешивания на базе преобразования Фурье, за которым следует подуровень прямой связи.

Такие модификации уменьшают сложность сети без потери качества, в сравнении с BERT и другими трансформерами, что позволяет достичь большей производительности (до 7 раз) [6] при обучении и работе сети. Указанные преимущества повышают перспективность использования трансформеров для задач компьютерного зрения на встраиваемых устройствах.

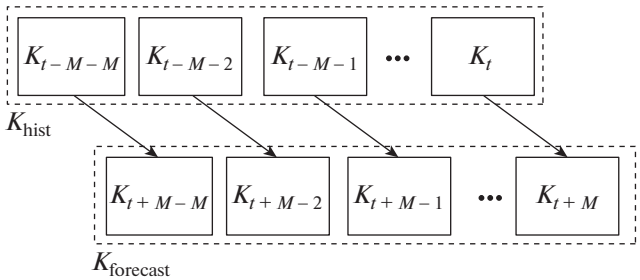

Рис. 4. Дискретный шаг прогнозирования.

Таблица 1. Результаты экспериментов

Прогнозируемый кадр	**30**				
Метрики	**R^2**	**MSE**	**RMSE**	**MAPE**	**MAE**
Базовая архитектура					
KeyLSTMv1	0.914254705	28.67433603	1002.186281	0.04443831	21.72306531
KeyFNetv1	0.951835	21.10453622	520.4710225	0.034808935	15.5945849
Разница KeyLSTMv1/KeyFNetv1	0.037580296	−7.569799813	−481.7152584	−0.009629375	−6.128480414
Модификация 1					
KeyLSTMv2	0.968188102	16.83534826	339.8362631	0.026397628	11.53717729
KeyFNetv2	0.969046236	16.69747483	334.6964651	0.026907618	11.77245659
Разница KeyLSTMv1/KeyFNetv2	0.000858133	−0.137873427	−5.139798001	0.000509989	0.235279305
Разница KeyFNetv1/KeyFNetv2	0.017211235	−4.407061388	−185.7745574	−0.007901317	−3.822128303
Модификация 2					
KeyFNetv3	0.972698527	15.6822523	293.3287594	0.024749771	10.89408894
Разница KeyFNetv2/KeyFNetv3	0.003652291	−1.015222533	−41.36770571	−0.002157847	−0.878367657
Прогнозируемый кадр			**60**		
Базовая архитектура					
KeyLSTMv1	0.89062067	32.61716063	1290.512357	0.051051181	23.43395894
KeyFNetv1	0.917591465	27.93320102	918.4545783	0.04548377	20.41898538
Разница KeyLSTMv1/KeyFNetv1	0.026970795	−4.683959614	−372.0577789	−0.005567411	−3.014973561
Модификация 1					
KeyLSTMv2	0.934810558	24.92314826	741.5463371	0.038525275	17.12124281
KeyFNetv2	0.939627948	24.04318982	692.2964439	0.038306814	16.86789884
Разница KeyLSTMv1/KeyFNetv2	0.00481739	−0.879958445	−49.24989322	−0.00021846	−0.253343976
Разница KeyFNetv1/ KeyFNetv2	0.022036484	−3.890011199	−226.1581345	−0.007176956	−3.551086544
Модификация 2					
KeyFNetv3	0.943740305	23.10762016	644.7951896	0.036441727	16.2068619
Разница KeyFNetv2/KeyFNetv3	0.004112357	−0.935569661	−47.50125431	−0.001865087	−0.661036941
Прогнозируемый кадр			**90**		
Базовая архитектура					
KeyLSTMv1	0.870611835	35.4546043	1498.300195	0.055390165	24.61419807
KeyFNetv1	0.886077554	33.28393733	1297.951204	0.052896528	24.20843555
Разница KeyLSTMv1/KeyFNetv1	0.015465719	−2.170666973	−200.3489908	−0.002493638	−0.405762515
Модификация 1					
KeyLSTMv2	0.906536961	30.16275159	1073.920934	0.047304529	20.9600912
KeyFNetv2	0.906031503	30.23749533	1080.924719	0.047517816	21.38840743
Разница KeyLSTMv1/KeyFNetv2	−0.000505458	0.074743742	7.003784273	0.000213287	0.428316234
Разница KeyFNetv1/KeyFNetv2	0.019953949	−3.046442004	−217.0264857	−0.005378712	−2.820028119
Модификация 2					
KeyFNetv3	0.912486804	29.25196562	1031.077395	0.045950611	20.39733995
Разница KeyFNetv2/KeyFNetv3	0.006455301	−0.985529711	−49.84732398	−0.001567205	−0.99106748
Прогнозируемый кадр			**120**		
Базовая архитектура					
KeyLSTMv1	0.851453504	38.28713612	1774.615252	0.057880437	26.65832968
KeyFNetv1	0.858961061	37.24500311	1632.172772	0.059275549	26.92123845
Разница KeyLSTMv1/KeyFNetv1	0.007507558	−1.042133007	−142.44248	0.001395113	0.262908771

Таблица 1. Окончание

Прогнозируемый кадр	30				
Метрики	**R^2**	**MSE**	**RMSE**	**MAPE**	**MAE**
Модификация 1					
KeyLSTMv2	0.88307481	33.87372826	1344.643725	0.053314063	23.88759837
KeyFNetv2	0.876368517	34.83015263	1454.909941	0.054228039	24.61600208
Разница KeyLSTMv1/KeyFNetv2	–0.006706293	0.956424368	110.2662152	0.000913976	0.72840371
Разница KeyFNetv1/KeyFNetv2	0.017407455	–2.414850483	–177.2628312	–0.00504751	–2.30523637
Модификация 2					
KeyFNetv3	0.883425931	33.83159518	1376.46009	0.051875596	23.41985738
Разница KeyFNetv2/KeyFNetv3	0.007057414	–0.998557447	–78.44985049	–0.002352443	–1.196144703

Блок преобразования Фурье применяет 2D Дискретное преобразование Фурье (DTF) к эмбеддингу входных данных: одно 1D DTF вдоль входных временных последовательностей F_{seq} и одно вдоль скрытой размерности слоя F_h (размерность эмбеддинга) [6]:

$$y = \Re(F_{seq}(F_h(x))). \quad (7)$$

В базовой архитектуре блок h_{ts} представлен в двух вариантах: или как слой Bi-LSTM, или как 2 последовательных блока FNet. Исходя из используемых блоков мы назвали архитектуры KeyLSTM, и KeyFNet.

В результате обучения обоих вариантов архитектур, KeyLSTM и KeyFNet соответственно, были получены следующие метрики на тестирующей подвыборке D^s_{valid} , представленные в табл. 1.

Результаты сравнительного анализа обученных архитектур глубоких нейронных сетей, в которых в качестве блока для выявления взаимосвязи во временных данных и координатах были использованы блоки LSTM и FNet, показывают, что использование блока FNet позволяет в среднем увеличить коэффициент детерминации на 0.0218 в сравнении с LSTM.

Был проведен анализ результатов нейросетевых блоков h_{ts} на различных временных диапазонах. Использование блока FNet по сравнению с блоком Bi-LSTM позволяет получить следующий прирост по метрике R^2:

• для целевой позиции на 30 кадре на 0.03758, при этом изменение по оси X увеличивается на 0.05178, для оси Y на 0.023381;

• для целевой позиции на 60 кадре на 0.026971 по метрике R^2, при этом изменение по оси X увеличивается на 0.044463, для оси Y на 0.009478;

• для целевой позиции на 90 кадре на 0.015466 по метрике R^2, при этом изменение по оси X увеличивается на 0.020568, для оси Y на 0.010363;

• для целевой позиции на 120 кадре на 0.007508 по метрике R^2, при этом изменение по оси X увеличивается на 0.0182, для оси Y уменьшается на 0.00318.

Также можно наблюдать, что увеличение временного диапазона прогноза целевой позиции человека снижает коэффициент детерминации для KeyFNet в среднем на 0.03095, а для KeyLSTM в среднем на 0.02093 по метрике R^2, из-за увеличения стохастичности поведения человека.

Модификация 1

Для улучшения результата предложенная архитектура была модифицирована путем добавления слоя h^{relu}_{fcn} для выявления взаимосвязей через кодирование координат x_n, y_n, где n – количество ключевых точек человека в соответствии с выбранным стандартом. Фактически координаты каждой точки были объединены путем конкатенации и переданы на полносвязную НС с функцией активации ReLU.

Модификация архитектуры отображена на рис. 3 соответствующим блоком.

В результате обучения обоих вариантов KeyLSTMv2 и KeyFNetv2 были получены метрики на тестирующей подвыборке $\mathbf{D}^s_{\text{valid}}$, представленные в табл. 1.

В зависимости от использования блока LSTM или блока FNet были получены следующие показатели увеличения метрики R^2: KeyFNetv2 для 30 кадров на 0.020864, для 60 кадров на 0.022036, для 90 кадров на 0.019954, для 120 кадров на 0.017407. KeyLSTMv2для 30 кадров на 0.053933, для 60 кадров на 0.04419, для 90 кадров на 0.035925, для 120 кадров на 0.031621.

Для отдельных осей координат коэффициент детерминации увеличился следующим образом:

• для целевой позиции на 30 кадре по оси X увеличился на 0.06613,, по оси Y увеличился на 0.038365 для KeyLSTMv2, по оси X увеличился на

0.015263, по оси Y увеличился на 0.019255 для KeyFNetv2;

• для целевой позиции на 60 кадре по оси X увеличился на 0.051177, по оси Y увеличился на 0.034737 для KeyLSTMv2, по оси X увеличился на 0.015263, по оси Y увеличился на 0.019255 для KeyFNetv2;

• для целевой позиции на 90 кадре по оси X увеличился на 0.040032, по оси Y увеличился на 0.029961 для KeyLSTMv2, по оси X увеличился на 0.015263, по оси Y увеличился на 0.019255 для KeyFNetv2;

• для целевой позиции на 120 кадре по оси X увеличился на 0.040796, по оси Y увеличился на 0.019377 для KeyLSTMv2, по оси X увеличился на 0.015263, по оси Y увеличился на 0.019255 для KeyFNetv2.

Предложенная модификация позволила достигнуть лучших показателей метрик (табл. 1), что свидетельствует о выявлении внутренней взаимосвязи значений координат.

Модификация 2

Результаты тестирования модифицированной архитектуры демонстрируют, что при использовании блоков FNet достигается значительное увеличение точности по сравнению с LSTM. Это обусловлено тем, что, фактически, последовательные блоки LSTM представляют собой 2 идентичных слоя реккурентно замкнутых на себя блока, в то время как трансформер – это широкая сеть с большим количеством параметров и механизмами внимания. Это позволяет трансформерам лучше "понимать" контекст и взаимосвязь отдельных признаков. В задаче прогнозирования движений человека под контекстом понимается взаимосвязь отдельных точек между собой, отражающих кинематику человеческого тела.

Таким образом, дальнейшими шагами исследования является улучшение точности при использовании только блоков FNet.

В качестве второй модификации предлагается преобразовать блок h_{ts} следующим образом:

1. На первый слой FNet передать вектор, представляющий собой параметры каждой точки после блока конкатенации.

2. Полученный вектор с первого слоя FNet передать на полносвязную нейронную сеть с функцией активации ReLU.

3. На следующий блок FNet передать выход из предыдущего слоя FNet и выход из полносвязного слоя.

4. Повторить шаги 2 и 3.

Итоговая модификация (рис. 3) в дальнейшем обозначается как KeyFNetv3.

В результате обучения KeyFNetv3 были получены результаты на тестирующей подвыборке $\mathbf{D}^{s}_{\text{valid}}$, приведенные в табл. 1.

Модификация KeyFNetv3 позволила увеличить коэффициент детерминации для 30 кадров на 0.003652, для 60 кадров на 0.004112, для 90 кадров на 0.006455, для 120 кадров на 0.007057.

Для отдельных осей координат коэффициент детерминации увеличился следующим образом по сравнению с KeyFNetv2:

• для целевой позиции на 30 кадре на 0.003652 при этом изменение по оси X увеличивается на 0.003593, для оси Y 0.003441;

• для целевой позиции на 60 кадре на 0.004112 по метрике R^2, при этом изменение по оси X увеличивается на 0.00445, для оси Y 0.00361;

• для целевой позиции на 90 кадре на 0.006455 по метрике R^2, при этом изменение по оси X увеличивается на 0.002929, для оси Y 0.01002;

• для целевой позиции на 120 кадре на 0.007057 по метрике R^2, при этом изменение по оси X увеличивается на 0.007563, для оси Y 0.00632.

Таким образом, предложенная итоговая модификация архитектуры KeyFNet показывает коэффициент детерминации 0.972699 и среднеквадратическую ошибку прогноза 293.328759 для оценки положения позиции человека на следующих 30 кадров.

Анализируя отдельные точки [12], можно наблюдать некоторое падение метрик на руках и локтях особенно при увеличении целевой прогнозной точки с 30 до 120 кадров. Так, например, разница по R2 для кисти и локтя между 30 и 120 кадром составляет до 0.05. Наблюдается корреляция между падением метрик для смежных точек (кисть, локоть, плечо), что позволяет сделать вывод, о том что алгоритм в целом уловил биомеханику человеческого тела и связь между частями тела, а падение метрик вызвано именно неправильной прогнозной траекторией.

При этом основные точки, которые сосредоточены вокруг центра человека (талия, шея, плечи), показывают высокий результат даже на 120 кадре. Это позволяет сделать вывод, что алгоритм правильно прогнозирует общую траекторию движения человека и стилистику походки.

Усреднение прогнозных значений с использованием дискретных временных шагов

Так как работа алгоритма подразумевает использование "сканирующего окна" с настраиваемым шагом, то возникает эффект перекрывающихся областей [4]. Для построения прогнозной траектории движения отдельных ключевых точек человека в видеопоследовательности предлагается использовать дискретный временной шаг для предсказания:

$$(f(\mathbf{K}_{\text{hist}_{t-M}}), f(\mathbf{K}_{\text{hist}_{t-1}}), \ldots, f(\mathbf{K}_{\text{hist}_t})) \rightarrow \rightarrow (\mathbf{K}_{t+M-M}, \mathbf{K}_{t+M-M-1}, \ldots \mathbf{K}_{t+M}), \quad (8)$$

где f – архитектура предложенной нейронной сети, $\mathbf{K}_{\text{hist}_t}$ – набор исторических кадров на момент времени t, $\mathbf{K}_t$ – положение координат.

Исходя из экспериментов (табл. 1) в качестве целевой точки предпочтительнее использовать оценку на 30 кадров вперед, что минимизирует ложно-негативные срабатывания и гарантирует более точное возможное положение для предотвращения травмоопасной ситуации. При этом для построения траектории движения используется комплексная оценка за счет каскадного подкрепления, в котором используется несколько моделей, поочередно формирующих траекторию движения человека с дискретным шагом.

4. ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Полунатурный эксперимент с собственными наборами данных

Предложенный подход реализован на языке python и протестирован на оборудовании с характеристиками CPU Core i9-12900 GPU Nvidia GeForce 4090.

Были собраны, размечены и аннотированы собственные наборы, содержащие видеозаписи рабочего процесса в центре робототехники. Съемка выполнялась с использованием мультикамерной системы, состоящей из трех камер: RealSense D455, RealSense L515 и RTCP CAM AC–D6144. Схема расположения устройств и их параметры приведены в работе [6].

Для разметки использовался инструмент MS COCO skeleton, а также разметка с использованием классификаторов OpenPifPaf [6] и MediaPipe [6]. Общая продолжительность размеченного видео составила 22 мин 17 с. Итоговое количество кадров составило 40 110.

На рис. 5 приведен пример прогнозирования на кадрах с камеры центра робототехники: а) текущий кадр, б) прогнозируемый кадр, где синие точки – прогнозируемое положение, красные точки – размеченное значение с использованием MediaPipe. При этом на текущей сцене достигаются следующие показатели по метрикам MSE: 22,1764; MAE: 22,1176; RMSE 1463,82352941; таким образом, отклонения отдельных точек не превышают 22 пикселя.

В результате расчетов на подвыборке реальных данных были получены следующие метрики с использованием итоговой архитектуры KeyFNet MSE: 31,9957; MAE: 29,7458; RMSE: 2172,9564.

Эксперимент с 3D

В рамках развития работы итоговая архитектура KeyFNetv3 была модифицирована в KeyFNetv3D для обработки трехмерных данных путем добавления Z координаты во входной вектор. В результате обучения были получены результаты на тестирующей подвыборке $\mathbf{D}^s_{\text{valid}}$. На различных временных диапазонах средний коэффициент детерминации достигает следующих значений: для 30 кадров 0.96503, для 60 кадров 0.91102, для 90 кадров 0.856697, для 120 кадров 0.802904.

В дальнейшем планируется переразметка собственного набора данных для получения пространственных глобальных 3D координат ключевых точек.

5. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТОВ

Проверка алгоритма осуществляется как на открытых контрольных наборах, так и на реальных данных с мультикамерной системы.

При использовании модифицированной архитектуры KeyFNetv3 для прогнозирования положения человека на будущих кадрах видеопотока на различных временных диапазонах средний коэффициент детерминации достигает следующих значений: для 30 кадров 0.972698527, для 60 кадров 0.943740305, для 90 кадров 0.912486804, для 120 кадров 0.883425931.

Применение дискретных шагов может быть использовано не только для построения прогнозных траекторий, но и для валидации и устранения выбросов, а также корректировки прогнозирования скорости движения объекта.

Эксперимент на полунатурных данных демонстрирует перспективность применения предложенного подхода для реальных задач коллаборативной робототехники.

Результаты применения предложенной архитектуры с 3D данными подтверждают, что предложенная архитектура обладает высокой степенью универсальности и может быть использована для любого стандарта ключевых точек, в том числе и для 3D представления.

Результаты доказывают, что предложенный алгоритм на базе трансформеров с модификацией имеет следующие преимущества перед классическими алгоритмами:

- повышение точности за счет выявления семантических отношений сложных пространственно-временных взаимодействий между частями тела;
- возможность адаптировать под 3D координаты без значимой потери качества;
- снижение вычислительных требований за счет использования блоков FNet на базе быстрого преобразования Фурье.

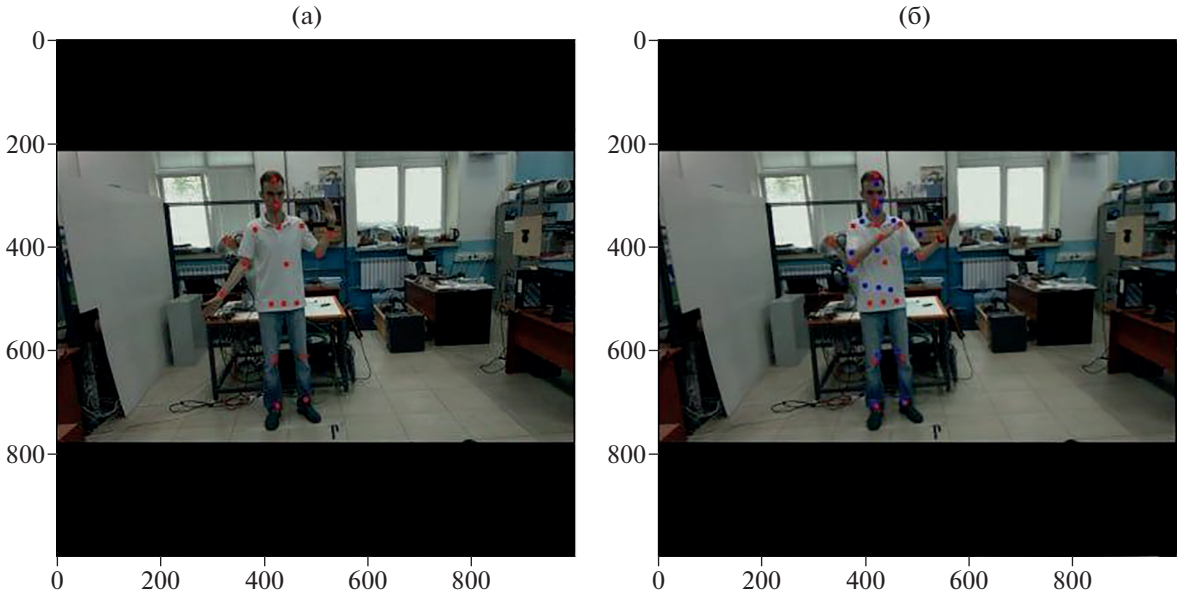

Рис. 5. Прогнозирование на кадрах с реальной камеры.

6. ЗАКЛЮЧЕНИЕ

В настоящей работе решена задача прогнозирования положения человека на будущих кадрах видеопотока с помощью трансформерной нейронной сети, которая соответствует по точности передовым моделям, но превосходит их по скорости и использует меньше ресурсов для вычисления.

Предложен ряд модификаций и разработана оригинальная архитектура глубокой нейронной сети для прогнозирования положения человека на будущих кадрах видеопотока.

Научная новизна состоит в применении комбинированного алгоритма на основе нескольких блоков FNet с быстрым преобразованием Фурье в качестве механизма внимания, конкатенирующих координаты ключевых точек.

Был проведен глубокий анализ различных способов построения архитектур для решения поставленной задачи. Проведены экспериментальные исследования по использованию как традиционных для данной задачи блоков LSTM, так и перспективных SOTA решений. Проведен сравнительный анализ с различным временным диапазоном прогноза позиции человека и различными блоками для выделения взаимосвязи во временных данных. Проведена комплексная оценка множественной регрессии. При использовании FNet достигается значительное увеличение точности по сравнению с LSTM.

Проведены полунатурные эксперименты с видеокадрами, полученными с мультикамерной системы центра робототехники, а также с 3D координатами. В результате тестирования на реальных данных, показатели точности модели остались рамках допустимых значений.

По сравнению с классическим подходом, предложенная архитектура показывает коэффициент детерминации 0.972699 и среднеквадратическую ошибку прогноза 293.328759 для оценки положения позиции человека на следующих 30 кадров.

Основными особенностями предложенной архитектуры являются: объединение и кодирование координат ключевых точек для нахождения закономерности в их расположении; использование нескольких блоков трансформерной модели FNet для выявления семантических отношений сложных пространственно-временных взаимодействий между частями тела, возможность быстрой адаптации под новый формат данных без существенных изменений архитектуры. Таким образом, указанные преимущества позволяют применять предложенный алгоритм в реальных системах промышленной и коллаборативной робототехники, что подтверждается результатами экспериментов.

Дальнейшими шагами исследования является улучшение точности за счет использования дополнительной аугментации, а также разметка расширенного набора данных и обучение с использованием дополнительных синтетических примеров.

В оптимизации предполагается выполнить квантование разработанной нейронной сети и перенести предложенный подход на аппаратную вычислительную базу Nvidia Jetson Nano.

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Работа выполнена при поддержке Российского научного фонда (проект № 22-71-10093 https://rscf.ru/project/22-71-10093/).

СПИСОК ЛИТЕРАТУРЫ

1. *Pintea S.L., van Gemert J.C., Smeulders A.W.M.* Déja vu: Motion prediction in static images // Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part III 13. Springer International Publishing, 2014. P. 172–187.
2. *Walker J., Gupta A., Hebert M.* Dense optical flow prediction from a static image // Proceedings of the IEEE International Conference on Computer Vision. 2015. P. 2443–2451.
3. *Chao Y.W. и др.* Forecasting human dynamics from static images // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. P. 548–556.
4. *Amosov O. и др.* Human localization in video frames using a growing neural gas algorithm and fuzzy inference // Comput. Opt. 2017. V. 41. № 1. P. 46–58. https://doi.org/10.18287/2412-6179-2017-41-1-46-58
5. *Amosov O.S. и др.* Using the deep neural networks for normal and abnormal situation recognition in the automatic access monitoring and control system of vehicles // Neural Comput. Appl. 2021. V. 33. № 8. P. 3069–3083. https://doi.org/10.1007/s00521-020-05170-5
6. *Gerasimenko N.A., Chernyavsky A.S., Nikiforova M.A.* RuSciBERT: A transformer language model for obtaining semantic embeddings of scientific texts in Russian // Dokl. Math. 2022. V. 106. № S1. P. S95–S96. https://doi.org/10.1134/S1064562422060072
7. *Amosov O.S. и др.* Using the ensemble of deep neural networks for normal and abnormal situations detection and recognition in the continuous video stream of the security system // Procedia Comput. Sci. 2019. V. 150. P. 532–539. https://doi.org/10.1016/j.procs.2019.02.089
8. *Gao X. и др.* Accurate grid keypoint learning for efficient video prediction // 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2021. P. 5908–5915. https://doi.org/10.1109/IROS51168.2021.9636874
9. *Liu Z. и др.* Swin transformer v2: Scaling up capacity and resolution // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. P. 12009–12019. https://doi.org/10.1109/CVPR52688.2022.01170
10. *Ionescu C. и др.* Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments // IEEE Trans. Pattern Anal. Mach. Intell. 2014. V. 36. № 7. P. 1325–1339. https://doi.org/10.1109/TPAMI.2013.248
11. *Ivanov Y. и др.* Using an ensemble of deep neural networks to detect human keypoints in the workspace of a collaborative robotic system // INTELS'22. Basel Switzerland: MDPI, 2023. https://doi.org/10.3390/engproc2023033019
12. GutHub : URL: https://github.com/IdentySergey/fnet (дата обращения: 25.08.2023)
13. *Lee-Thorp J. и др.* Fnet: Mixing tokens with fourier transforms //arXiv preprint arXiv:2105.03824. 2021.
14. *Kreiss S., Bertoni L., Alahi A.* Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association //IEEE Transactions on Intelligent Transportation Systems. 2021. V. 23. № 8. P. 13498–13511. https://doi.org/10.1109/tits.2021.3124981
15. *Lugaresi и др.* Mediapipe: A framework for building perception pipelines //arXiv preprint arXiv:1906.08172. 2019.

INVESTIGATION OF NEURAL NETWORK ALGORITHMS FOR HUMAN MOVEMENT PREDICTION BASED ON LSTM AND TRANSFORMERS

S. V. Zhiganov[a], Y. S. Ivanov[a], and D. M. Grabar[a]

[a] *Komsomolsk-on-Amur State University, Komsomolsk-on-Amur, Russian Federation*

Presented by Academician of the RAS A.I. Avetisyan

Experiments on Human3.6M and on our own real data confirmed the effectiveness of the proposed approach based on FNet blocks, compared to the traditional approach based on LSTM. The proposed algorithm matches the accuracy of advanced models, but outperforms them in terms of speed and uses less computational resources and can be applied in collaborative robotic solutions. The problem of predicting the position of a person on future frames of a video stream is solved and in-depth experimental studies on the application of traditional and SOTA blocks for this task are carried out. An original architecture of KeyFNet and its modifications based on transform blocks is presented, which is able to predict coordinates in the video stream for 30, 60, 90 and 120 frames ahead with high accuracy. The novelty lies in the application of a combined algorithm based on multiple FNet blocks with fast Fourier transform as an attention mechanism concatenating the coordinates of key points.

Keywords: prediction key points, transformers, collaborative robotic systems, deep learning