

doi: 10.17586/2226-1494-2026-26-5-1039-1048

УДК 004.932.2

## Метод автоматического распознавания коммуникативных жестов участников виртуального взаимодействия

Анастасия Александровна Двойникова<sup>1✉</sup>, Алексей Анатольевич Карпов<sup>2</sup>

<sup>1,2</sup> Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация

<sup>1</sup> dvoynikova.a@ias.spb.su✉, <https://orcid.org/0000-0001-8047-6639>

<sup>2</sup> karpov@ias.spb.su, <https://orcid.org/0000-0003-3424-652X>

### Аннотация

**Введение.** Межличностное общение все чаще происходит в виртуальном формате с помощью систем телеконференций. Несмотря на преимущества такого формата, ограниченный размер видеоклика, нестабильность интернет-соединения и снижение качества передачи информации затрудняют восприятие невербальных сигналов участников коммуникации. В процессе межличностного взаимодействия коммуникативные жесты являются важным средством передачи эмоций и вовлеченности, а их автоматический анализ способен повысить точность систем распознавания поведенческих и психологических состояний участников виртуальной коммуникации. В работе предлагается иерархический двухуровневый метод распознавания коммуникативных жестов в условиях виртуальной коммуникации, позволяющий выполнять совместное распознавание статических и динамических жестов. **Метод.** Иерархия метода заключается в последовательном использовании моделей для распознавания типа жеста, а затем классификации статических и динамических жестов. На первом этапе выполняется определение типа жеста: статический, динамический или отсутствие жеста. На втором этапе используются специализированные модели для детального распознавания трех классов статических и трех классов динамических жестов. Для анализа видеопотока исследуются различные типы признаков, включая ключевые точки тела, их производные геометрические признаки, оптический поток и признаки предобученной модели Vision Transformer (ViT). В качестве классификаторов рассматриваются полносвязные, сверточно-рекуррентные и трансформерные нейронные сети. Дополнительно исследуются методы повышения точности классификации при дисбалансе классов и малом объеме данных. **Основные результаты.** Экспериментальные исследования проводились на многомодальном корпусе данных ENERGI, содержащем данные групповой виртуальной коммуникации с аннотацией коммуникативных жестов. Установлено, что наилучшие результаты для распознавания статических жестов достигаются при использовании признаков ViT и сверточно-рекуррентной нейронной сети, где средняя F-мера составила 70,4 %. Для динамических жестов наилучшей оказалась комбинация производных геометрических признаков и полносвязной нейронной сети со средней F-мерой 58,1 %. Для определения типа жестов наиболее эффективным оказалось совместное использование признаков ViT и оптического потока, обеспечившее среднюю F-меру 76,5 %. Итоговая средняя F-мера распознавания 7 классов коммуникативных жестов составила 68,7 %, тогда как для одноуровневого метода — 20,8 %. Также показано, что применение методов компенсации дисбаланса классов позволяет дополнительно повысить качество классификации. **Обсуждение.** Полученные результаты показывают эффективность иерархического подхода по сравнению с одноуровневой классификацией благодаря разделению задачи на этапы определения типа жеста и последующего специализированного распознавания. Установлено, что различные типы признаков наиболее эффективны для разных категорий жестов: признаки ViT лучше описывают статические конфигурации позы, тогда как производные геометрические признаки более информативны для анализа движений. Предложенный метод может использоваться в интеллектуальных системах анализа поведения участников виртуальной коммуникации, включая задачи оценивания вовлеченности, эмоционального состояния и эффективности взаимодействия.

### Ключевые слова

коммуникативные жесты, статические жесты, динамические жесты, иерархический метод, нейронные сети, дисбаланс классов

© Двойникова А.А., Карпов А.А., 2026

**Благодарности**

Работа выполнена в рамках бюджетной темы Санкт-Петербургского Федерального исследовательского центра РАН (№ FFZF-2025-0003).

**Ссылка для цитирования:** Двойникова А.А., Карпов А.А. Метод автоматического распознавания коммуникативных жестов участников виртуального взаимодействия // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 5. С. 1039–1048. doi: 10.17586/2226-1494-2026-26-5-1039-1048

## Automatic recognition method for communicative gestures of virtual interaction participants

Anastasia A. Dvoynikova<sup>1</sup>✉, Alexey A. Karpov<sup>2</sup>

<sup>1,2</sup> St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation

<sup>1</sup> dvoynikova.a@ias.spb.su✉, <https://orcid.org/0000-0001-8047-6639>

<sup>2</sup> karpov@ias.spb.su, <https://orcid.org/0000-0003-3424-652X>

**Abstract**

Currently, interpersonal communication increasingly takes place in virtual environments through teleconferencing systems. Despite the advantages of this format, limited video window size, unstable Internet connections, and reduced transmission quality hinder the perception of nonverbal signals exchanged by participants. During interpersonal interaction, communicative gestures serve as an important means of conveying emotions and engagement, while their automatic analysis can improve the accuracy of systems designed to recognize behavioral and psychological states of participants in virtual communication. This paper proposes a hierarchical method for communicative gesture recognition in virtual communication environments, enabling the joint recognition of both static and dynamic gestures. The proposed method is based on the sequential application of models for gesture type recognition followed by the classification of static and dynamic gestures. At the first stage, gesture type is determined as static, dynamic, or no gesture. At the second stage, specialized models are employed to recognize three classes of static gestures and three classes of dynamic gestures. Various feature types are investigated for video analysis, including body keypoints, their geometric derivative features, optical flow, and features extracted from a pretrained Vision Transformer (ViT) model. Fully connected, convolutional-recurrent, and transformer-based neural networks are considered as classifiers. In addition, methods for improving classification accuracy under class imbalance and limited data conditions are investigated. Experimental studies were conducted using the multimodal ENERGI dataset containing recordings of group virtual communication annotated with communicative gestures. The best performance for static gesture recognition was achieved using ViT features and a convolutional-recurrent neural network, resulting in an average F1-score of 70.4 %. For dynamic gesture recognition, the most effective combination consisted of geometric derivative features and a fully connected neural network, achieving an average F1-score of 58.1 %. For gesture type recognition, the joint use of ViT features and optical flow yielded the highest average F1-score of 76.5 %. The overall average F1-score for recognizing seven communicative gesture classes reached 68.7 %, whereas the single-model approach achieved only 20.8 %. The results also demonstrate that applying class imbalance compensation techniques further improves classification performance. The results obtained confirm the effectiveness of the hierarchical approach compared to single-model classification due to the decomposition of the task into gesture type recognition and subsequent specialized classification. Different feature types were found to be most effective for different gesture categories: ViT features better capture static pose configurations, whereas geometric derivative features provide more informative representations of motion patterns. The proposed method can be applied in intelligent systems for behavioral analysis of participants in virtual communication, including engagement assessment, emotion recognition, and interaction effectiveness evaluation.

**Keywords**

communicative gestures, static gestures, dynamic gestures, hierarchical method, neural networks, class imbalance

**Acknowledgements**

This work was prepared within the framework of the budget topic of the St. Petersburg Federal Research Center of the Russian Academy of Sciences (No. FFZF-2025-0003).

**For citation:** Dvoynikova A.A., Karpov A.A. Automatic recognition method for communicative gestures of virtual interaction participants. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 5, pp. 1039–1048 (in Russian). doi: 10.17586/2226-1494-2026-26-5-1039-1048

**Введение**

В настоящее время межличностное общение все чаще происходит в виртуальном формате с помощью систем телеконференций [1–3]. Несмотря на очевидные преимущества такого формата, он также имеет ряд ограничений. Они связаны со сложностью понимания невербальных сигналов, которые зависят от чувств,

эмоций и внутренних состояний участников. Эти трудности могут быть обусловлены ограниченным размером видеоокна, нестабильностью интернет-соединения, а также сниженным качеством передачи информации по каналам связи, что приводит к потере части значимых поведенческих и коммуникативных признаков [3].

Анализ виртуальной коммуникации приобретает особую значимость в образовательной, профессио-

нальной и социальной сферах. Понимание поведенческих и невербальных сигналов участников позволяет более точно оценивать их состояние, вовлеченность и эффективность взаимодействия [3]. По этой причине в настоящее время задача автоматического анализа подобных данных с использованием методов машинного обучения является актуальной [4].

В процессе межличностного взаимодействия люди используют коммуникативные жесты для передачи эмоций и вовлеченности [5]. В условиях онлайн-коммуникации набор используемых жестов сокращается из-за ограниченного размера кадра, однако движения рук и головы остаются важными индикаторами вовлеченности и эмоций участников. Отметим, что автоматический анализ коммуникативных жестов способен повысить точность систем распознавания поведенческих и психологических состояний участников виртуальной коммуникации [6].

### Обзор существующих методов

На сегодняшний день существует немного исследований по анализу коммуникативных жестов [1, 2, 7–11]. В работах [7, 8] выполнен анализ наличия и распознавания динамических жестов, а в [9] предложена система распознавания статических и динамических жестов. В работе [7] рассмотрена задача классификации четырех типов коммуникативных жестов в повествовательном взаимодействии. При использовании признаков, включающих характеристики движения и контекстную информацию, в [7] получено значение точности 75,6 %. В [8] осуществлено распознавание 12 классов жестов с использованием координат составных углов в качестве признаков с точностью 98,6 %.

Также существуют мультимодальные подходы, использующие аудио, текст, позы тела, мимику и жесты рук [9, 10]. Например, в работе [9] предложен мультимодальный набор данных из 6 классов жестов, на котором модель Long Short-Term Memory обеспечивает точность (Accuracy) 92 %; при этом набор данных сформирован на основе записей одного испытуемого. В [10] предложена мультимодальная система распознавания 6 классов прагматических динамических жестов, для которой достигнуто значение взвешенной F1-меры 62,5 %.

Отдельное внимание уделяется распознаванию жестов в условиях виртуального взаимодействия. Так, в работе [1] предложена система анализа четырех классов коммуникативных жестов верхней части тела, для которой достигнуто значение средней F-меры 50,4 %. В [2] представлена система классификации 30 классов жестов, связанных с взаимодействием рук и лица, достигающей Accuracy 88,7 %.

Современные тенденции использования больших языковых моделей также находят применение в данной области. В работе [11] модель ChatGPT используется для распознавания трех классов демонстрационных жестов диктора, которая достигает средней F-меры 44,7 %.

Большинство методов анализа жестов включают этапы сегментации, извлечения признаков и класси-

фикации [6]. Сегментация (Region of Interest) — выделение области изображения, представляющей интерес для анализа (чаще всего — рука, голова или тело). В качестве методов сегментации применяют метод Continuously Adaptive Mean Shift [6], Histogram of Oriented Gradients [6], Виолы–Джонса [1].

Для извлечения визуальной информации из изображений с жестами часто используются признаки: ключевые точки рук в 2D- [7, 9] и 3D-проекции [9, 10], ключевые точки лица и тела [1], производные ключевых точек (например, расстояния и углы между точками) [8], а также нейросетевые признаки, такие как TResNet [2].

Методы классификации жестов варьируются в зависимости от наличия динамики в анализируемых жестах. Для классификации статических жестов применяются различные традиционные методы [6], а также нейросетевые модели [6], такие как Vision Transformer (ViT) [2]. В качестве классификаторов динамических жестов используются скрытые марковские модели [7], метод опорных векторов [7, 8], рекуррентные нейросетевые модели [9, 10], а также модель ChatGPT [11].

Проведенный обзор показывает, что существующие исследования в области анализа коммуникативных жестов в основном ориентированы на решение узкоориентированных задач, таких как бинарное определение наличия жеста [7] или распознавание ограниченного числа классов [1, 2]. Кроме того, большинство работ рассматривает только статические [1, 2] или динамические жесты [8–10], что ограничивает применение таких подходов в условиях естественной коммуникации. Также следует отметить, что существующие исследования [1, 2, 7–10], как правило, используют закрытые наборы данных, что затрудняет воспроизводимость результатов и прямое сравнение различных методов. В отличие от рассмотренных подходов, в настоящей работе предлагается единая система распознавания статических и динамических жестов на данных виртуальной коммуникации.

### Исследовательский корпус данных

Для экспериментальных исследований использован многомодальный корпус данных ENgagement and Emotion Russian Gathering Interlocutors (ENERGI) [3]. Корпус содержит 6,4 ч аудиовизуальных данных участников групповой виртуальной коммуникации. Данные аннотированы по классам вовлеченности, активации и валентности эмоций, а также 11 классам коммуникативных жестов. Аннотация в корпусе выполнена для каждого участника коммуникации по 10-секундным непересекающимся интервалам (окнам). Корпус данных разделен на поднаборы по принципу дикторонезависимости в соотношении 70/15/15 % на обучающий, валидационный и тестовый поднаборы. Данные собраны в естественных условиях взаимодействия, из-за чего классы жестов несбалансированы и наблюдаются несоответствия их распределений между поднаборами. Для устранения несоответствия часть миноритарных классов («Рука на носу», «Зевота», «Рука закрывает глаза», «Скрещенные руки») была объединена в класс 1 «Нет жестов». Итоговое распределение классов ком-

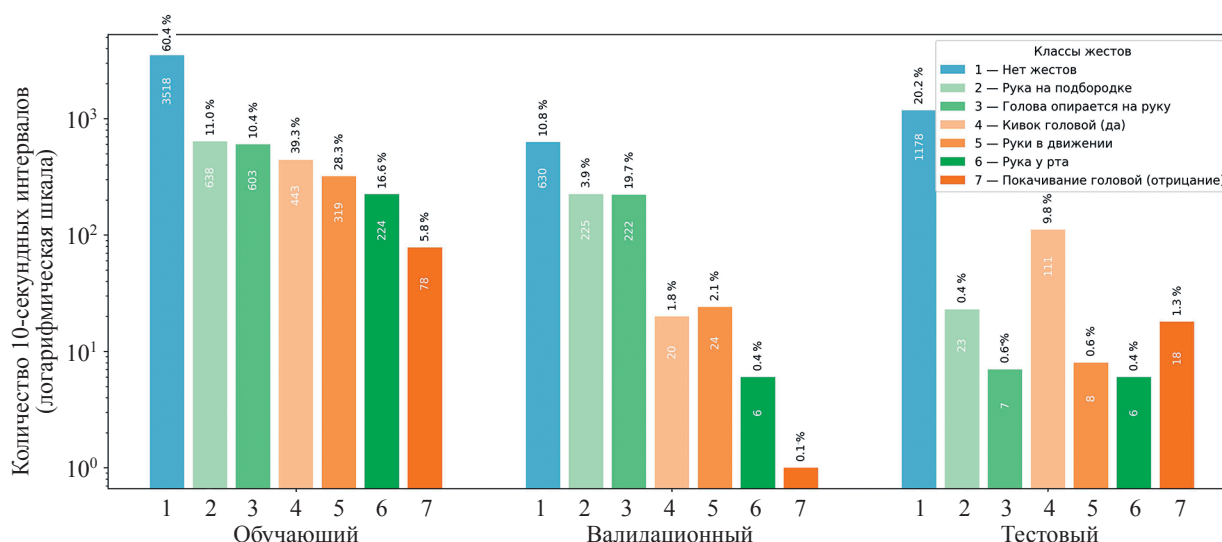


Рис. 1. Распределение данных в классах коммуникативных жестов по трем поднаборам

Fig. 1. Distribution of data across communicative gesture classes

коммуникативных жестов по поднаборам представлено на рис. 1.

Как видно из рис. 1, классы коммуникативных жестов несбалансированы. В корпусе ENERGI представлены статические (классы 2, 3 и 6) и динамические (классы 4, 5 и 7) жесты, при этом отсутствие жестов (класс 1) также является информативным признаком вовлеченности, поэтому класс 1 рассматривается как значимый.

### Метод распознавания коммуникативных жестов

Исходя из особенностей данных корпуса ENERGI разработан иерархический двухуровневый метод распознавания коммуникативных жестов (рис. 2).

Иерархия предложенного метода заключается в последовательном использовании нескольких моделей для распознавания коммуникативных жестов. На первом этапе выполняется классификация по типу жеста на классы статических, динамических или отсутствию жестов. На втором этапе для статических и динамических классов обучается собственная модель, которая распознает по три класса жестов.

Перед обучением моделей из исходных 10-секундных видео извлекаются кадры. Дальнейшая обработка выполняется на каждом втором кадре. В качестве признаков используются координаты ключевых точек тела и их производные геометрические признаки (ПГ), оптический поток (ОП) и признаки ViT [12].

Ключевые точки извлекаются с помощью архитектуры Keypoint R-CNN [13] после масштабирования до  $256 \times 256$  пикселей и перевода в формат RGB. Используются 11 точек верхней части тела, на основе которых вычисляются расстояния, углы, а также их первая и вторая производные (скорость и ускорение движения). Для ОП и предобученного на базе данных аннотированных изображений ImageNet [14] ViT применяется детектирование лица RetinaFace [15]: лицо переводится в градацию серого и масштабируется до

$128 \times 128$  пикселей для ОП, а для ViT — до  $224 \times 224$  пикселей в формате RGB.

### Экспериментальные исследования иерархического метода

Экспериментальные исследования проводились с различными видеопризнаками и моделями: полносвязной (ПС) и сверточно-рекуррентной нейронными (СРН) сетями, а также на основе трансформеров для каждой подзадачи в иерархическом методе распознавания. Архитектуры ПС и СРН сетей показаны на рис. 2. Архитектура нейронной сети на основе трансформера состоит из линейного слоя, позиционного кодирования, двух последовательных блоков-трансформеров и линейных слоев. Для каждой сети и каждой подзадачи выполнялся подбор оптимальных гиперпараметров. В процессе подбора рассматривались различные значения размера батча (8, 16 и 32), а также варианты функции потерь: кросс-энтропия и фокальная как с использованием весов классов, так и без них. Выбор оптимальной конфигурации осуществлялся на основе значения средней F-меры, полученной на валидационной выборке. Оптимальные гиперпараметры для всех сетей следующие: размер батча — 16; 30 эпох с ранней остановкой при отсутствии увеличения точности на 5 эпохах; шаг обучения —  $10^{-4}$ ; фокальная функция потерь; оптимизатор — AdamW и планировщик обучения — ReduceLROnPlateau. В качестве интегрального показателя оценки точности работы сетей использовалась средняя F-мера. В табл. 1 представлены результаты экспериментальных исследований иерархического метода распознавания коммуникативных жестов. Лучшая комбинация (рис. 2) представляет собой модели, использующие взвешивающую конкатенацию ОП + СРН + ViT + ПС для классификации типа жестов, ViT + СРН — статических жестов, ПГ + ПС — динамических жестов.

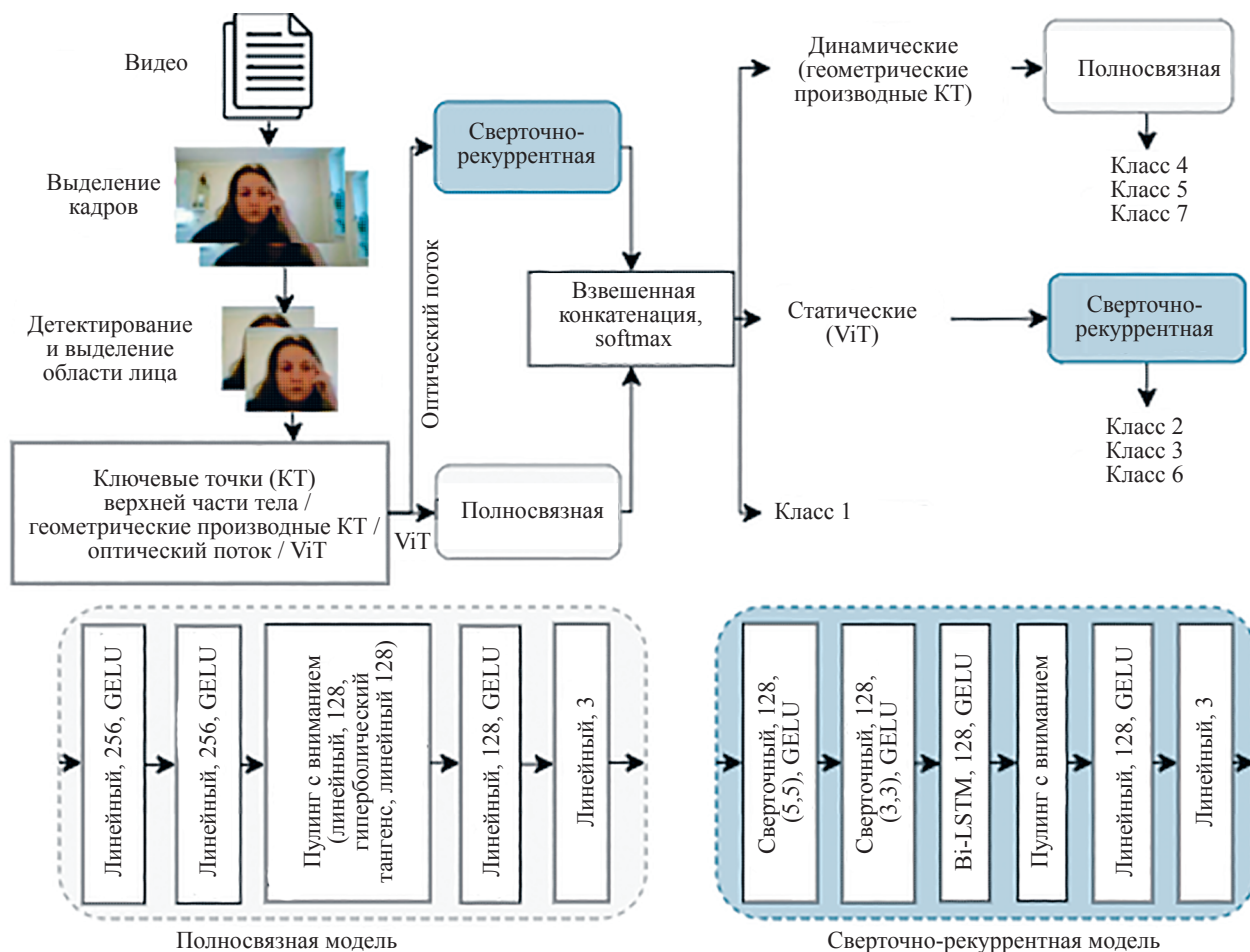


Рис. 2. Иерархический метод распознавания коммуникативных жестов.

GELU — функция активации Gaussian Error Linear Unit

Fig. 2. Hierarchical method for communicative gesture recognition

ческих жестов. Лучшая комбинация определялась по средней F-мере для каждой модели.

Как видно из табл. 1, наилучшей комбинацией признаков и моделей для распознавания статических жестов является ViT + CPN, а для динамических жестов — ПГ + ПС. Из результатов экспериментов можно увидеть, что для модели распознавания типа жеста лучшие результаты показывают несколько комбинаций. ViT + ПС, которые распознают с наибольшей точностью статические жесты (73,5 %) и их отсутствие (91,1 %). При этом комбинация ОП и CPN сети показывают наивысший результат по распознаванию класса динамических жестов (55,7 %). Для улучшения результата первой модели иерархической системы были проведены эксперименты по объединению двух лучших комбинаций, которые позволили получить уверенные результаты для каждого класса и наибольший по показателю средней F-меры 76,5 %. Наилучшая комбинация методов представлена на рис. 1.

Для оценки преимущества использования иерархического метода распознавания коммуникативных жестов выполнены экспериментальные исследования с базовым одноуровневым методом распознавания одновременно всех 7 классов коммуникативных жестов. Были проведены аналогичные эксперименты со всеми

визуальными признаками и рассматриваемыми моделями для одноуровневого метода. Лучшей комбинацией признаков и моделей для распознавания одновременно 7 классов являются признаки ViT и нейронная сеть трансформер. Выполнено сравнение иерархического метода с результатами одноуровневого метода распознавания 7 классов жестов (табл. 2).

Из табл. 2 видно, что использование иерархического подхода позволяют получить среднюю F-меру для 7 классов равную 68,7 %. При этом одноуровневый метод показывает F-меру 20,8 %, что на 47,9 % ниже иерархического.

#### Экспериментальные исследования методов повышения точности классификации при дисбалансе классов и малого объема данных

Как видно из анализа корпуса данных и результатов экспериментальных исследований, для повышения точности распознавания необходимо применять дополнительные методы по регулированию дисбаланса классов и малого объема данных в классах. В настоящей работе исследовались методы случайной взвешенной выборки, фокальная функция потерь и взвешивание классов в ней, а также температурное масштабирование. Метод

Таблица 1. Результаты экспериментальных исследований иерархического метода распознавания коммуникативных жестов по F-мере, %

Table 1. Experimental results of the hierarchical communicative gesture recognition method: the F-score, %

| Признаки                                     | Модель распознавания типа жестов |             |              |             | Модель статических жестов |             |             |             | Модель динамических жестов |             |             |             |
|----------------------------------------------|----------------------------------|-------------|--------------|-------------|---------------------------|-------------|-------------|-------------|----------------------------|-------------|-------------|-------------|
|                                              | класс 1                          | статические | динамические | среднее     | класс 2                   | класс 3     | класс 6     | среднее     | класс 4                    | класс 5     | класс 7     | среднее     |
| Полносвязная модель                          |                                  |             |              |             |                           |             |             |             |                            |             |             |             |
| ключевые точки (КТ)                          | 63,5                             | 4,8         | 25,3         | 31,2        | 55,6                      | 36,4        | 16,0        | 36,0        | 67,0                       | 34,3        | 13,3        | 38,2        |
| ПГ                                           | 84,5                             | 14,2        | 10,4         | 36,4        | 29,4                      | 36,4        | 12,5        | 26,1        | 80,2                       | 48,0        | <b>46,2</b> | <b>58,1</b> |
| ОП                                           | 72,1                             | 5,6         | 32,4         | 36,7        | 40,0                      | 33,3        | 44,4        | 39,3        | 60,4                       | 33,3        | 30,8        | 41,5        |
| ViT                                          | 91,1                             | 73,5        | 30,0         | 64,9        | <b>86,8</b>               | 36,4        | 28,6        | 50,0        | 26,0                       | 18,2        | 21,2        | 21,8        |
| Сверточно-рекуррентная модель                |                                  |             |              |             |                           |             |             |             |                            |             |             |             |
| КТ                                           | 77,9                             | 11,0        | 12,8         | 33,9        | 14,3                      | 66,7        | 25,0        | 35,3        | 75,4                       | 36,4        | 18,9        | 43,5        |
| ПГ                                           | 77,6                             | 12,2        | 3,5          | 31,1        | 19,4                      | 25,0        | 16,0        | 20,1        | <b>87,7</b>                | <b>50,0</b> | 32,3        | 56,6        |
| ОП                                           | 75,7                             | 0,96        | <b>55,7</b>  | 44,1        | 22,2                      | 24,2        | 66,7        | 37,7        | 60,4                       | 33,3        | 30,8        | 41,5        |
| ViT                                          | 72,3                             | 53,1        | 31,2         | 52,3        | 80,0                      | <b>86,8</b> | 44,4        | <b>70,4</b> | 63,5                       | 0,0         | 19,1        | 27,5        |
| Трансформерная модель                        |                                  |             |              |             |                           |             |             |             |                            |             |             |             |
| КТ                                           | 64,9                             | 10,9        | 24,1         | 33,3        | 85,2                      | 20,0        | 0,0         | 35,1        | 59,1                       | 30,0        | 12,9        | 34,0        |
| ПГ                                           | 78,1                             | 9,7         | 11,1         | 32,9        | 68,2                      | 25,0        | 20,0        | 37,8        | 26,2                       | 11,7        | 14,3        | 17,4        |
| ОП                                           | 31,1                             | 1,9         | 25,9         | 19,6        | 63,6                      | 11,7        | <b>72,7</b> | 49,4        | 70,9                       | 26,1        | 39,5        | 45,5        |
| ViT                                          | 78,5                             | 60,8        | 32,4         | 57,1        | 85,2                      | 18,2        | 28,6        | 44,0        | 75,0                       | 0,0         | 30,8        | 35,3        |
| Лучшая комбинация                            |                                  |             |              |             |                           |             |             |             |                            |             |             |             |
| Модели, использующие взвешивающую комбинацию | ОП + CPN + ViT + ПС              |             |              |             | ViT + CPN                 |             |             |             | ПГ + ПС                    |             |             |             |
|                                              | <b>95,2</b>                      | <b>75,8</b> | <b>58,7</b>  | <b>76,5</b> | 80,0                      | <b>86,8</b> | 44,4        | <b>70,4</b> | 80,2                       | 48,0        | <b>46,2</b> | <b>58,1</b> |

Примечание: жирным шрифтом выделены лучшие результаты.

случайной взвешенной выборки позволяет увеличить вероятность выбора миноритарных классов при формировании обучающих батчей. Взвешивание классов заключается в назначении больших весов миноритарным классам при вычислении функции потерь. Фокальная функция потерь дополнительно уменьшает влияние легко классифицируемых примеров и концентрирует обучение на сложных и редких классах. Температурное масштабирование использовалось для сглаживания распределения логитов модели и снижения уверенности классификатора при обучении на несбалансированных данных. Результаты экспериментальных исследований с различными методами повышения точности классификации при дисбалансе классов и их комбинаций

между собой представлены в табл. 3, показывающей результаты с использованием иерархического метода с лучшими признаками и классификаторами как без регулирования дисбаланса классов, так и с ними, а также изменение результатов ( $\Delta$ ).

### Обсуждение

Результаты экспериментальных исследований показывают, что наилучшие значения F-меры для модели распознавания типа жеста были получены при совместном использовании признаков ViT и ОП. Это может быть связано с тем, что признаки ViT позволяют эффективно учитывать пространственный контекст и конфи-

Таблица 2. Сравнение результатов иерархического и одноуровневого методов по F-мере, %

Table 2. Comparison of the results of hierarchical and single-model methods

| Метод         | Класс 1     | Модель статических жестов |             |         |             | Модель динамических жестов |         |             |             | Среднее по семи классам |
|---------------|-------------|---------------------------|-------------|---------|-------------|----------------------------|---------|-------------|-------------|-------------------------|
|               |             | класс 2                   | класс 3     | класс 6 | среднее     | класс 4                    | класс 5 | класс 7     | среднее     |                         |
| Иерархический | <b>95,2</b> | 80,0                      | <b>86,8</b> | 44,4    | <b>70,4</b> | 80,2                       | 48,0    | <b>46,2</b> | <b>58,1</b> | <b>68,7</b>             |
| Одноуровневый | 86,6        | 23,8                      | 0,0         | 20,0    | 14,6        | 1,6                        | 6,1     | 7,4         | 5,1         | 20,8                    |

Таблица 3. Результаты экспериментальных исследований методов повышения точности классификации при дисбалансе классов и малом объеме данных

Table 3. Experimental results of classification accuracy improvement methods for imbalanced classes and limited data

| Метод                                                                        | Модель распознавания типа жестов |          | Модель статических жестов |          | Модель динамических жестов |          |
|------------------------------------------------------------------------------|----------------------------------|----------|---------------------------|----------|----------------------------|----------|
|                                                                              | F-мера, %                        | $\Delta$ | F-мера, %                 | $\Delta$ | F-мера, %                  | $\Delta$ |
| Без методов устранения дисбаланса                                            | 73,6                             | 0,0      | 49,6                      | 0,0      | 34,8                       | 0,0      |
| Случайная взвешенная выборка                                                 | 74,2                             | 0,6      | 57,7                      | 8,1      | 39,1                       | 4,3      |
| Взвешивание классов                                                          | 68,4                             | -5,2     | 43,6                      | -6,0     | 34,5                       | -0,3     |
| Фокальная функция потерь                                                     | 71,4                             | -2,2     | 55,8                      | 6,2      | 29,7                       | -5,1     |
| Температурное масштабирование                                                | 74,7                             | 1,1      | 43,6                      | -6,0     | 48,8                       | 14,0     |
| Взвешивание классов, фокальная функция потерь                                | 74,0                             | 0,4      | <b>70,4</b>               | 20,8     | <b>58,1</b>                | 23,3     |
| Взвешивание классов, фокальная функция потерь, температурное масштабирование | <b>76,5</b>                      | 2,9      | 67,5                      | 17,9     | 51,7                       | 16,9     |

гурацию позы для распознавания статических жестов и отсутствия жестов, тогда как ОП более точно описывает движение и временную динамику. При этом ПГ показали наилучшие результаты для распознавания динамических жестов при использовании ПС сети (F-мера 58,1 %), поскольку такие признаки напрямую описывают траектории и взаимное расположение ключевых точек во времени. Для статических жестов наблюдается иная закономерность: наибольшую эффективность демонстрируют модели, использующие ViT-признаки. Это связано с тем, что статические жесты зависят от тонких конфигураций позы и контекста положения рук относительно тела, которые лучше кодируются глобальными визуальными представлениями.

Сравнение архитектур классификаторов показывает, что ПС сеть в большинстве случаев обеспечивает более высокие и стабильные результаты по сравнению со

СРН сетью и трансформером. Вероятно, это связано с ограниченным объемом и сильной несбалансированностью данных, при которых более сложные архитектуры склонны к переобучению и хуже обобщают признаки. Дополнительно наблюдается значительное различие в качестве распознавания отдельных классов жестов. Это может объясняться меньшим числом примеров данных классов и высокой визуальной схожестью с другими жестами или нейтральным состоянием.

Для детального анализа ошибок модели была построена нормализованная матрица спутывания для иерархического метода с наилучшей конфигурацией признаков и классификаторов (рис. 3).

Анализ нормализованной матрицы спутывания позволяет выявить ключевые закономерности ошибок модели. Наиболее уверенно распознается класс 1 (Ассигасу 97,2 %), при этом основные ошибки для

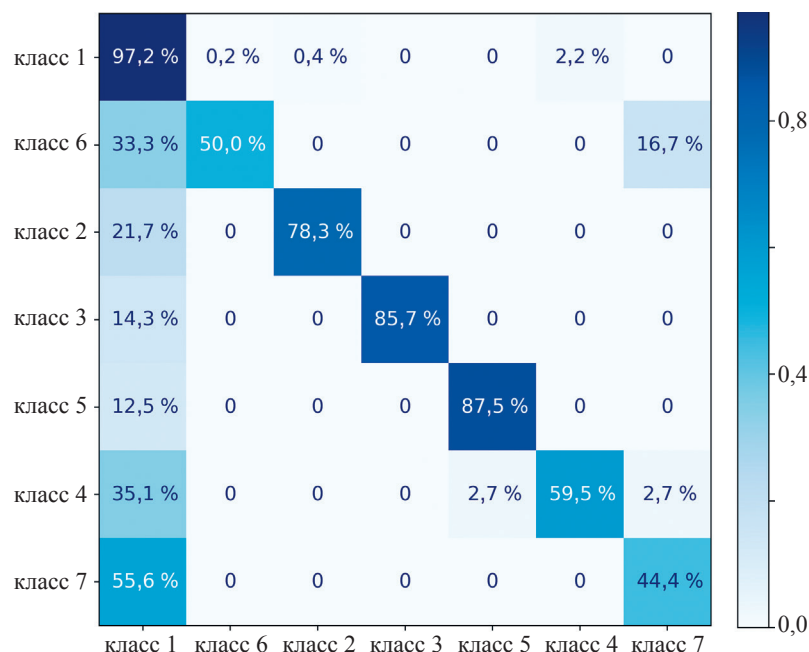


Рис. 3. Нормализованная матрица спутывания для иерархического метода распознавания коммуникативных жестов

Fig. 3. A normalized confusion matrix for the hierarchical communicative gesture recognition method

него приходится на класс 4 (2,2 %), что объясняется естественными микродвижениями головы в состоянии покоя. Жесты класса 2 (78,3 %), класса 3 (85,7 %) и класса 5 (87,5 %) демонстрируют высокую точность. Наибольшие сложности вызывает класс 6 (50,0 %), который в 33,3 % случаев ошибочно классифицируется как класс 1, что связано с нечеткой границей между данным жестом и естественным положением руки вблизи лица. Среди динамических жестов класс 4 распознается с точностью 59,5 %, при этом 35,1 % примеров ошибочно относятся к классу 1, что указывает на проблему различения малых кивков и естественных движений головы. Класс 7 показывает точность 44,4 % и в 55,6 % случаев путается с классом 1, вероятно, из-за недостаточной выраженности боковых движений в некоторых видеофрагментах. Важно отметить, что перекрестные ошибки между статическими и динамическими классами полностью отсутствуют, что подтверждает корректную работу первой ступени иерархического метода и эффективность разделения типов жестов.

Использование иерархического подхода позволило существенно повысить качество распознавания по сравнению с одноуровневым методом. Полученные результаты свидетельствуют о том, что разделение задачи на этапы определения типа жеста и последующей специализированной классификации позволяет снизить межклассовую неоднородность и уменьшить влияние дисбаланса данных.

Прямое сравнение полученных результатов с другими известными исследованиями затруднено. Это связано с тем, что большинство работ [1, 2, 7–11] использует собственные закрытые наборы данных и различные классификации коммуникативных жестов. Кроме того, многие исследования рассматривают только статические или динамические жесты. Предложенный метод объединяет оба типа жестов в рамках единой системы. При этом в опубликованных исследованиях значения средней F-меры варьируются от 44,7 до 62,5 % для различных постановок задачи, тогда как предложенный подход обеспечивает F-меру 68,7 %, что свидетельствует о его конкурентоспособности.

Результаты экспериментов по регулированию дисбаланса классов показывают, что эффективность методов зависит от типа модели и решаемой задачи. Применение метода случайной взвешенной выборки позволило повысить среднюю F-меру для всех моделей, тогда как использование только взвешивания классов в большинстве случаев привело к снижению качества распознавания. Фокальная функция потерь обеспечила прирост средней F-меры для моделей статических жестов на 6,2 %. Температурное масштабирование показывает значительное увеличение значения F-меры для модели динамических жестов на 14 %. Наилучшие результаты распознавания статических и динамических жестов были достигнуты при совместном использовании взвешивания классов и фокальной функции потерь, где средняя F-мера составила 70,4 % и 58,1 %, соответственно. Для модели определения типа жестов максимальное качество было получено при дополни-

тельном применении температурного масштабирования совместно с взвешиванием классов и фокальной функции потерь, что позволило повысить среднюю F-меру до 76,5 %. Полученные результаты подтверждают необходимость индивидуального подбора методов балансировки классов для каждой модели.

## Заключение

Представлен иерархический двухуровневый метод распознавания коммуникативных жестов в условиях групповой виртуальной коммуникации. Для экспериментальных исследований использовался собранный корпус данных ENERGI, содержащий статические и динамические жесты, а также характеризующийся сильной несбалансированностью классов и малым объемом данных. Предложенный подход включает два этапа: начальное определение типа жеста (статический, динамический или отсутствие жеста); последующее распознавание конкретного класса жеста. Выполнено исследование различных видеопризнаков, включая ключевые точки, производные геометрические признаки, оптический поток и Vision Transformer (ViT), а также полносвязные, сверточно-рекуррентные и трансформерные архитектуры классификаторов. Эксперименты показали, что наилучшие результаты для распознавания статических жестов достигаются при использовании признаков ViT и сверточно-рекуррентной сети (F-мера 70,4 %), а для динамических жестов — при использовании производных геометрических признаков и полносвязной нейронной сети (F-мера 58,1 %). Для определения типа жестов наиболее эффективным оказалось совместное использование признаков ViT и оптического потока, что позволило получить среднюю F-меру 76,5 %. Показано, что применение иерархического подхода существенно превосходит одномодельный метод распознавания 7 классов жестов, где средняя F-мера составила 68,7 % против 20,8 %. Дополнительно исследовались методы компенсации дисбаланса классов. Эксперименты показали, что эффективность таких методов зависит от решаемой задачи и используемой модели. Наилучшие результаты для распознавания статических и динамических жестов были получены при совместном использовании взвешивания классов и фокальной функции потерь, а для модели определения типа жестов — при дополнительном применении температурного масштабирования. Прямое сравнение результатов с существующими исследованиями затруднено из-за различий в используемых корпусах данных и наборах классов жестов. В отличие от большинства существующих работ, ориентированных отдельно на статические или динамические жесты, в данной работе предложен единый иерархический подход для их совместного распознавания в условиях виртуальной коммуникации. Полученные результаты могут быть использованы при разработке интеллектуальных систем анализа поведения участников виртуальной коммуникации. В частности, при анализе вовлеченности участников в групповую коммуникацию на основе анализа эмоций и коммуникативных жестов.

## Литература

1. Marcos-Ramiro A., Pizarro D., Barrios A., Donoser M. Body communicative cue extraction for conversational analysis // *Proc. of the 10<sup>th</sup> IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 2013. P. 1–8. doi: 10.1109/FG.2013.6553741
2. Roy K., Mohanty A., Ranjan Sahay R. Hand over face gesture classification with feature driven vision transformer and supervised contrastive learning // *Lecture Notes in Computer Science*. 2024. V. 15304. P. 400–417. doi: 10.1007/978-3-031-78128-5\_26
3. Двойникова А.А., Величко А.Н., Карпов А.А. Многомодальный корпус данных взаимодействия участников виртуальной коммуникации ENERGI // *Известия высших учебных заведений. Приборостроение*. 2025. Т. 68. № 12. С. 1011–1019. doi: 10.17586/0021-3454-2025-68-12-1011-1019
4. Kapitanov A., Kvanchiani K., Nagaev A., Kraynov R., Makhliarchuk A. HaGRID — HAND Gesture Recognition Image Dataset // *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2024. P. 4572–4581. doi: 10.1109/WACV57701.2024.00451
5. Wang T., Lin X., Xu Y., Ye Q., Guo D., Escalera S., et al. Micro-gesture recognition: A comprehensive survey of datasets, methods, and challenges // *Machine Intelligence Research*. 2026. V. 23. N 2. P. 308–330. doi: 10.1007/s11633-025-1629-x
6. Jalayer R., Jalayer M., Orsenigo C., Tomizuka M. A review on deep learning for vision-based hand detection, hand segmentation and hand gesture recognition in human–robot interaction // *Robotics and Computer-Integrated Manufacturing*. 2026. V. 97. P. 103110. doi: 10.1016/j.rcim.2025.103110
7. Okada S., Bono M., Takanashi K., Sumi Y., Nitta K. Context-based conversational hand gesture classification in narrative interaction // *Proc. of the 15<sup>th</sup> ACM on International conference on multimodal interaction*. 2013. P. 303–310. doi: 10.1145/2522848.2522898
8. Ishrak S., Munir M.B., Kabir M.H. Dynamic hand gesture recognition using sequence of human joint relative angles // *Proc. of the International Conference on Electrical, Computer and Communication Engineering (ECCE)*. 2023. P. 1–6. doi: 10.1109/ECCE57851.2023.10101606
9. Schak M., Gepperth A. Gesture recognition on a new multi-modal hand gesture dataset // *Proc. of the 11<sup>th</sup> International Conference on Pattern Recognition Applications and Methods — ICPRAM*. 2022. P. 122–131. doi: 10.5220/0010982200003122
10. Zhang J., Li S., Mawlim C.O., Okada S. Multimodal classification of co-speech gesture pragmatic function in storytelling // *Proc. of the 13<sup>th</sup> International Conference on Human-Agent Interaction*. 2025. P. 312–320. doi: 10.1145/3765766.3765771
11. Hensel L.B., Yongsatianchot N., Torshizi P., Minucci E., Marsella S. Large language models in textual analysis for gesture selection // *Proc. of the 25<sup>th</sup> International Conference on Multimodal Interaction*. 2023. P. 378–387. doi: 10.1145/3577190.3614158
12. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., et al. An image is worth 16×16 words: Transformers for image recognition at scale // *arXiv*. 2020. arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929
13. He K., Gkioxari G., Dollár P., Girshick R. Mask R-CNN // *Proc. of the IEEE International Conference on Computer Vision (ICCV)*. 2017. P. 2980–2988. doi: 10.1109/iccv.2017.322
14. Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L. ImageNet: A large-scale hierarchical image database // *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 2009. P. 248–255. doi: 10.1109/CVPR.2009.5206848
15. Deng J., Guo J., Ververas E., Kotsia I., Zafeiriou S. RetinaFace: Single-shot multi-level face localisation in the wild // *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. P. 5202–5211. doi: 10.1109/cvpr42600.2020.00525

## Авторы

Двойникова Анастасия Александровна — младший научный сотрудник, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, <https://orcid.org/0000-0001-8047-6639>, [dvoynikova.a@ias.spb.su](mailto:dvoynikova.a@ias.spb.su)

## References

1. Marcos-Ramiro A., Pizarro D., Barrios A., Donoser M. Body communicative cue extraction for conversational analysis. *Proc. of the 10<sup>th</sup> IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8. doi: 10.1109/FG.2013.6553741
2. Roy K., Mohanty A., Ranjan Sahay R. Hand over face gesture classification with feature driven vision transformer and supervised contrastive learning. *Lecture Notes in Computer Science*, 2024, vol. 15304, pp. 400–417. doi: 10.1007/978-3-031-78128-5\_26
3. Dvoynikova A.A., Velichko A.N., Karpov A.A. ENERGI: a multimodal data corpus of interaction of participants in virtual communication. *Journal of Instrument Engineering*, 2025, vol. 68, no. 12, pp. 1011–1019. (in Russian). doi: 10.17586/0021-3454-2025-68-12-1011-1019
4. Kapitanov A., Kvanchiani K., Nagaev A., Kraynov R., Makhliarchuk A. HaGRID — HAND Gesture Recognition Image Dataset. *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 4572–4581. doi: 10.1109/WACV57701.2024.00451
5. Wang T., Lin X., Xu Y., Ye Q., Guo D., Escalera S., et al. Micro-gesture recognition: A comprehensive survey of datasets, methods, and challenges. *Machine Intelligence Research*, 2026, vol. 23, no. 2, pp. 308–330. doi: 10.1007/s11633-025-1629-x
6. Jalayer R., Jalayer M., Orsenigo C., Tomizuka M. A review on deep learning for vision-based hand detection, hand segmentation and hand gesture recognition in human–robot interaction. *Robotics and Computer-Integrated Manufacturing*, 2026, vol. 97, pp. 103110. doi: 10.1016/j.rcim.2025.103110
7. Okada S., Bono M., Takanashi K., Sumi Y., Nitta K. Context-based conversational hand gesture classification in narrative interaction. *Proc. of the 15<sup>th</sup> ACM on International conference on multimodal interaction*, 2013, pp. 303–310. doi: 10.1145/2522848.2522898
8. Ishrak S., Munir M.B., Kabir M.H. Dynamic hand gesture recognition using sequence of human joint relative angles. *Proc. of the International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2023, pp. 1–6. doi: 10.1109/ECCE57851.2023.10101606
9. Schak M., Gepperth A. Gesture recognition on a new multi-modal hand gesture dataset. *Proc. of the 11<sup>th</sup> International Conference on Pattern Recognition Applications and Methods — ICPRAM*, 2022, pp. 122–131. doi: 10.5220/0010982200003122
10. Zhang J., Li S., Mawlim C.O., Okada S. Multimodal classification of co-speech gesture pragmatic function in storytelling. *Proc. of the 13<sup>th</sup> International Conference on Human-Agent Interaction*, 2025, pp. 312–320. doi: 10.1145/3765766.3765771
11. Hensel L.B., Yongsatianchot N., Torshizi P., Minucci E., Marsella S. Large language models in textual analysis for gesture selection. *Proc. of the 25<sup>th</sup> International Conference on Multimodal Interaction*, 2023, pp. 378–387. doi: 10.1145/3577190.3614158
12. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929
13. He K., Gkioxari G., Dollár P., Girshick R. Mask R-CNN. *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988. doi: 10.1109/iccv.2017.322
14. Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L. ImageNet: A large-scale hierarchical image database. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848
15. Deng J., Guo J., Ververas E., Kotsia I., Zafeiriou S. RetinaFace: Single-shot multi-level face localisation in the wild. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5202–5211. doi: 10.1109/cvpr42600.2020.00525

## Authors

Anastasia A. Dvoynikova — Junior Researcher, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, <https://orcid.org/0000-0001-8047-6639>, [dvoynikova.a@ias.spb.su](mailto:dvoynikova.a@ias.spb.su)

**Карпов Алексей Анатольевич** — доктор технических наук, профессор, руководитель лаборатории, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербург, 199178, Российская Федерация, [sc 57219469958](https://orcid.org/0000-0003-3424-652X), <https://orcid.org/0000-0003-3424-652X>, [karpov@iias.spb.su](mailto:karpov@iias.spb.su)

**Alexey A. Karpov** — D.Sc., Professor, Head of Laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint Petersburg, 199178, Russian Federation, [sc 57219469958](https://orcid.org/0000-0003-3424-652X), <https://orcid.org/0000-0003-3424-652X>, [karpov@iias.spb.su](mailto:karpov@iias.spb.su)

*Статья поступила в редакцию 01.06.2026*  
*Одобрена после рецензирования 31.08.2026*  
*Принята к печати 23.09.2026*

*Received 01.06.2026*  
*Approved after reviewing 31.08.2026*  
*Accepted 23.09.2026*



Работа доступна по лицензии  
Creative Commons  
«Attribution-NonCommercial»