УДК 004.93'12
doi:10.31799/1684-8853-2021-6-10-20

Научные обзоры
Surveys

Аналитический обзор моделей и методов автоматического распознавания жестов и жестовых языков

Д. А. Рюмин[a], *канд. техн. наук, старший научный сотрудник, orcid.org/0000-0002-7935-0569*
И. А. Кагиров[a], *научный сотрудник, orcid.org/0000-0003-1196-1117, kagirov@iias.spb.su*
А. А. Аксёнов[a], *младший научный сотрудник, orcid.org/0000-0002-7479-2851*
А. А. Карпов[a], *доктор техн. наук, доцент, главный научный сотрудник, orcid.org/0000-0003-3424-652X*
[a]*Санкт-Петербургский Федеральный исследовательский центр РАН, 14-я линия В. О., 39, Санкт-Петербург, 199178, РФ*

Введение: *распознавание жестов и жестовых языков является одной из наиболее развивающихся областей компьютерного зрения и прикладной лингвистики. Результаты исследований автоматического распознавания жестов находят применение в самых разных областях — от сурдоперевода до жестовых интерфейсов. В связи с этим разрабатываются разнообразные системы и методы для анализа жестовых данных.* ***Цель:*** *выполнить подробный обзор методов и провести сравнительный анализ существующих подходов в области автоматического распознавания жестов и жестовых языков.* ***Результаты:*** *анализ известных публикаций показал, что основными проблемами в области распознавания жестов являются детектирование артикуляторов (в первую очередь рук), распознавание их конфигурации и сегментация жестов в потоке речи. Сформулирован вывод о перспективности применения двухпоточных сверточных и рекуррентных архитектур нейросетей для эффективного извлечения и обработки пространственных и темпоральных признаков жеста, а также для решения проблемы автоматического распознавания динамических жестов и коартикуляций в потоке речи. При этом решение указанной проблемы напрямую зависит от качества и доступности наборов данных.* ***Практическая значимость:*** *представленный обзор рассматривается как вклад в изучение быстро развивающихся подходов к решению задачи распознавания жестовых данных независимо от материалов конкретных естественных жестовых языков. Результаты работы могут быть использованы при проектировании программных систем для автоматического распознавания жестов и жестовых языков.*

Ключевые слова *— жестовые языки, жестовый корпус, распознавание жестов, геометрия рук, машинное обучение, компьютерное зрение, нейросетевая модель, цифровая обработка, морфологическая обработка, информативные признаки.*

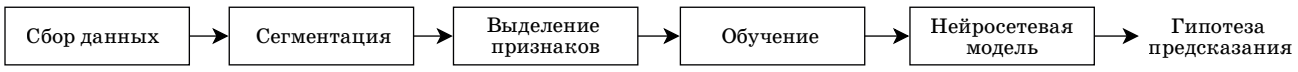
Для цитирования: Рюмин Д. А., Кагиров И. А., Аксёнов А. А., Карпов А. А. Аналитический обзор моделей и методов автоматического распознавания жестов и жестовых языков. *Информационно-управляющие системы*, 2021, № 6, с. 10–20. doi:10.31799/1684-8853-2021-6-10-20

For citation: Ryumin D. A., Kagirov I. A., Axyonov A. A., Karpov A. A. Analytical review of models and methods for automatic recognition of gestures and sign languages. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2021, no. 6, pp. 10–20 (In Russian). doi:10.31799/1684-8853-2021-6-10-20

Введение

На сегодня в сфере автоматического распознавания жестов и компьютерного зрения применяется широкий круг методов; несколько упрощая ситуацию, но нисколько ее не искажая, можно констатировать, что чаще всего автоматическое распознавание жеста складывается из этапов, представленных на рис. 1.

На первом этапе создается жестовый корпус, необходимый для обучения нейросетевой модели. Зачастую одновременно происходит предобработка (нормализация) входных данных (изображений или полноцветных (RGB) кадров видеопотока), при которой используется широкий круг методов цифровой обработки: яркостное преобразование; сглаживание и повышение резкости; масштабирование; пространственная фильтрация и т. д. Следующий этап представлен морфологической обработкой (дилатация, эрозия, обнаружение перепадов и др.) и сегментацией изображений на отдельные области интереса. После предобработки становится возможным извлечение информативных признаков, приме-



■ ***Рис. 1.*** Этапы извлечения информативных признаков жеста и создания модели на основе нейросетей
■ ***Fig. 1.*** Stages of informative features extraction and creation of a neural network-based model

няемых на последующем этапе для обучения нейросети и создания нейросетевой модели. В современных интеллектуальных системах признаки, как правило, извлекаются самой нейросетью и не выносятся в отдельный этап.

Следует обратить внимание на то, что в представленной схеме (см. рис. 1) отсутствует этап, на котором происходит собственно перевод на/с жестового языка. Такое решение обусловлено, прежде всего, тем, что системы, ориентированные на перевод, ограничиваются в основном лабораторными проектами или предназначены для перевода звучащего языка (зачастую — его текстового представления) в жестовую форму (Zardoz, TEAM, ASL Synthesizer, «Сурдофон»). Кроме того, для анализа высказываний на жестовом языке предварительно требуется система, с высокой точностью идентифицирующая жесты, и именно распознаванию жестов посвящено подавляющее большинство публикаций по теме, вышедших в последнее время [1, 2].

Разберем основные методы и подходы, применяемые на этапах, отраженных на рис. 1.

Жестовые наборы данных и корпусы

Как правило, любое исследование в области машинного обучения начинается с использования какого-либо из существующих наборов данных или корпусов (отличается наличием аннотации и сегментации) или со сбора нового. В последнем случае исследователи собирают набор данных и аннотируют его, основываясь на поставленной цели и определенных задачах. Широкое распространение получили корпусы, содержащие отдельные слова, числительные и буквы. В настоящее время в открытом доступе существует множество жестовых наборов данных и корпусов, собранных разными исследователями для разных целей (табл. 1).

Все представленные выше корпусы различаются количеством жестов, аппаратными средствами видеозахвата, фоновой обстановкой и, самое главное, задачами, для которых они создавались. Большинство корпусов в свободном доступе, существующих на сегодня, собраны и аннотированы для задач распознавания конфигураций рук диктора или отдельных жестов; корпусов, предназначенных для распознавания слитной жестовой речи, гораздо меньше. Как правило, для подобных исследований принято собирать собственный корпус. Также следует отметить, что единственный относительно крупный аннотированный корпус для русского жестового языка предназначен для решения образовательных и (или) строго лингвистических задач (http://rsl.nstu.ru/).

Сегментация изображения

Под сегментацией изображений понимается разбиение исходного изображения на непохожие по ряду признаков области, при этом предполагается, что каждая такая область содержит отдельный объект или его часть, а границы между областями соответствуют границам между объектами или их частями на исходном изображении. В случае с задачей автоматического распознавания жестовой информации основной целью сегментации изображения является выделение графических областей интереса, содержащих артикуляторы диктора (в первую очередь руки).

■ ***Таблица 1***. Крупнейшие доступные жестовые корпусы (с 2014 г.)
■ ***Table 1***. The largest open-source sign corpora (since 2014)

Ссылка	Год	ЗР	ПО	Классы	Записи	Р	УЗ
[3]	2014	Слитная жестовая речь	9	1200	45 760	+	ЦК
[4]	2016	Изолированные жесты	21	249	47 933	+	Kinect
[5]	2017	Конфигурация и ориентация руки	н/д	н/д	1,7 млн кадров	–	Искусственная генерация
[6]			10	496	2,2 млн кадров	+	Intel RealSense
[7]	2018	Изолированные жесты	50	83	2081	+	
[8]			9	1066	67 781	–	ЦК
[9]	2020		10	100	10 000	+	
[10]		Конфигурация	26	93	2,6 млн кадров	–	
[11]		Слитная жестовая речь	н/д	2000	21 083	+	
[12]				226	38 336	+	Kinect

Примечание: ЗР — задача распознавания; ПО — количество предметных областей; Р — разметка; УЗ — устройство захвата (камера, сенсор); ЦК — цветная камера; н/д — нет данных.

Сегментация изображений может проводиться вручную, полуавтоматически и автоматически. Результаты ручной сегментации попрежнему считаются лучшими по качеству [13], однако ручная сегментация не всегда возможна при работе с большими наборами данных, поскольку требует больших человеческих и временных затрат. Полуавтоматические методы сегментации частично решают эту проблему, однако попрежнему требуют человеческого присутствия. Последнее особенно важно в контексте сегментации, поскольку экспертная разметка все же не свободна от ошибок; кроме того, разметка одного и того же набора данных может разниться у разных экспертов. В то же время единожды разработанный алгоритм автоматической сегментации изображения всегда будет работать одинаково.

В литературе выделяются разнообразные методы автоматической сегментации изображений (корреляционные, пороговые и текстурные методы, глубокое обучение и т. п.); следует отметить, что одним из самых старых методов распознавания жестов является определение артикуляторов по цвету кожи [14].

Современные методы сегментации подразумевают использование машинного обучения. Большинство из них основано на семействе архитектур R-CNN. В принципе, нейросетевая модель R-CNN (и ее модификации Fast R-CNN и Faster R-CNN [15, 16]) является частным случаем техники автоматической сегментации изображения; к основным недостаткам R-CNN можно отнести энергозатратность (продолжительное время на процесс обучения) и неспособность функционировать в режиме реального времени.

R-CNN и развивающие ее архитектуры CNN (Convolutional Neural Network, 'сверточная нейронная сеть') не работают со всем изображением, подаваемым на вход. Для решения этой проблемы была разработана другая архитектура CNN, а именно You Only Look Once (YOLO) [17]. YOLO разбивает входное изображение на сетку (матрицу пикселей), в дальнейшем обрабатывая каждый сегмент отдельно. Классификация каждого сегмента производится с помощью единственной CNN. Недостаток YOLO заключается в невысокой эффективности сегментации и классификации мелких объектов на изображении. В работе [18] была предложена другая CNN обнаружения объектов (Single Shot Multi-box Detector), которая работает при разных ориентациях искомых объектов, устойчива к окклюзиям, а также работает в режиме реального времени.

В большинстве современных нейросетевых моделей для обнаружения областей рук используются архитектуры CNN или же их комбинации с другими разновидностями нейронных сетей, например долгая кратковременная память (Long Short-Term Memory — LSTM) [19, 20], которая способна извлекать пространственно-временные признаки жеста из последовательностей ранее сегментированных областей интереса.

Извлечение информативных визуальных признаков

Этап извлечения информативных признаков является ключевым в любой интеллектуальной системе, предназначенной для автоматического распознавания жестов. Извлеченные признаки прямо влияют на конечный результат распознавания, поскольку именно на этом этапе создаются входные данные для последующей классификации и образуется гипотеза предсказания.

Современные методы извлечения признаков в контексте распознавания жестовых языков можно разбить на следующие группы.

1. Методы, основанные на внешнем виде (appearance-based) [21, 22], подразумевают извлечение визуальных признаков для построения модели артикулятора (внешнего вида). Иногда признаки вычисляются на основании степени интенсивности пикселей, без предварительной сегментации. Такие методы работают в реальном времени благодаря довольно простому процессу извлечения признаков из двумерных (2D) данных. В работе [22] извлечение признаков основано на примитивах Хаара. Преимущество этого подхода состоит в способности к анализу контрастности между темными и яркими объектами на изображении и низким уровнем зависимости от окклюзий и динамики освещения.

2. N-точечные модели [23] и детектирование по карте глубины (3D) [24] характерны для тех случаев, когда основным устройством захвата видеоинформации является сенсор с датчиком глубины, такой как Azure Kinect или Leap Motion Controller. N-точечные модели работают с признаками скелета руки, задающими ее геометрические характеристики (например, ориентацию пальцев и расстояние между ними).

3. Методы на основе 3D-моделей используют трехмерный видеопоток информации в виде карты глубины или трехмерного облака точек до элементов артикулятора, фактически позволяя оперировать объемной моделью руки. Для обучения таких нейросетевых моделей хорошо подходят архитектуры 3D CNN [25]. Существенный недостаток архитектур 3D CNN заключается в том, что для их обучения требуется наличие довольно большого корпуса данных.

Среди прочих методов извлечения признаков можно упомянуть анализ главных компонент, линейный дискриминантный анализ и метод опорных векторов [26].

Обучение нейросетевой модели

Следует констатировать, что на сегодня преобладающим направлением в области автоматического распознавания жестов (и, в отдельных случаях, собственно распознавания жестового языка [27]) являются методы машинного обучения. Применение таких методов, в частности моделей CNN и LSTM, для обозначенной задачи обусловлено их высоким потенциалом извлечения как пространственных, так и временных признаков из видеопотока.

Можно выделить основные задачи, для решения которых применяются методы машинного обучения: 1) оценка конфигурации артикулятора; 2) распознавание жестов в потоке слитной речи.

Оценка конфигурации и ориентации руки артикулятора

Оценка положения руки — это процесс моделирования руки в виде набора ее некоторых частей (например, ладони и пальцев) для определения их положения в пространстве. Существуют подходы к решению этой задачи (например, [28]), оперирующие фалангами пальцев, однако почти во всех современных работах рука представляется по суставам пальцев, что сводит задачу моделирования фактически к нахождению положения всех суставов.

Чаще всего методы оценки конфигурации кистей рук подразумевают вычисление вероятности положения отдельных областей рук, соответствующих фалангам, пальцам и ладоням, на основе анализа 2D-плоскости полноцветного изображения или моделирования скелета кисти в 3D-пространстве. Так, например, в работе [29] карта глубины использовалась для оценки положения каждого из 21 элемента кисти руки диктора, однако локализация областей, содержащих кисти рук, производилась на изображениях в формате RGB.

Основная трудность 2D-подходов состоит в том, что уменьшение размера входных данных с 3D до 2D значительно усложняет задачу. Обыкновенно для обучения сети с использованием изображений в формате RGB требуется гораздо больше данных, чем для обучения аналогичной сети с использованием карт глубины [30–32].

К современным методам анализа 3D-информации относится применение генеративно-состязательных нейросетей (Generative Adversarial Network — GAN) [33], которые состоят из двух взаимосвязанных нейросетевых моделей (генератора и дискриминатора). Цель генератора — создавать новые изображения, а дискриминатора — оценивать их подлинность. В упомянутой работе использовался алгоритм CyclicGAN [34], причем генератор порождал конфигурации рук, основываясь на 3D-представлении кисти диктора. Другой распространенной техникой является моделирование псевдотрехмерной (2.5D) [35] руки на основе данных, полученных от карты глубины.

■ ***Таблица 2***. Оценка конфигурации кисти при помощи нейросетевых моделей

■ ***Table 2***. Hand pose estimation with the use of deep-learning models

Ссылка	Основная архитектура	Тип данных	Набор данных	Средняя мера ошибки
[35]	CNN	3D	MSRA Hand [41]	8
[37]	Вариационный автоэнкодер	RGB+3D	ICVL [42]	19,5
[33]	GAN	3D		8,5
[38]	CNN	3D		6,28
[6]	CNN	RGB+3D	BigHand2.2M [10]	17,1
[33]	GAN	3D		13,7
[39]	CNN	2D (RGB)	STB [43]	8,34
[40]	CNN	RGB+3D		5

Для сравнения в табл. 2 представлены основные архитектуры нейросетевых моделей и их значения ошибки при формировании гипотезы предсказания относительно конфигурации кисти руки, сгруппированные по двум тестовым наборам данных из корпусов BigHand2.2M и ICVL [36].

Распознавание жестовой информации в потоке слитной речи

Перечисленные выше методы достаточно хорошо работают в том случае, если речь идет о распознавании изолированных и статических жестов. Ситуация усложняется, если жест оказывается динамическим и (или) включенным в цепочку жестов, иными словами, при распознавании слитной жестовой речи. Основным отличием данной задачи от задачи распознавания статических жестов является наличие временных признаков и зачастую отсутствие предварительной разметки (т. е. ассоциации последовательностей кадров с жестами). Другую проблему, связанную с распознаванием слитной речи, представляет так называемая эпентеза — межжестовые движения, возникающие в потоке речи [44]. Появление больших корпусов в свободном доступе (начиная с RWTHPHOENIX-Weather-2014 и других из табл. 1) сделало возможным применение методов

■ ***Таблица 3.*** Современные нейросетевые модели, применяющиеся для распознавания слитной жестовой речи
■ ***Table 3.*** Current NN-based models for continuous sign language recognition

Ссылка	Тип нейросетевой модели	Архитектура	Характеристики
[45]	CNN+LSTM+ скрытая марковская модель	Google-LeNet	Использование нескольких модальностей: обучение на размеченных изображениях руки и рта
[46]	3D CNN+Bi-LSTM	LS-HAN	Двухпоточная сеть: одновременно обрабатываются область руки и весь кадр в целом
[47]	3D CNN+LSTM	I3D	Использование псевдометок
[48]	3D CNN, Bi-LSTM	VGG-S	Поэтапная настройка модели за счет признаков, извлеченных CNN. Использование нескольких модальностей (RGB и оптический поток)
[49]	3D CNN	3D-ResNet	Использование классификаторов лемм и *n*-граммов для определения слов в высказывании

машинного обучения для распознавания слитной жестовой речи. Сегодня основным способом обработки жестов в потоке речи является применение различных архитектур CNN и LSTM (табл. 3).

Ряд нейросетевых моделей из табл. 3 позволяют работать с несколькими модальностями. Так, в [50] описана иерархическая двухпоточная архитектура CNN, предназначенная для обнаружения и классификации жестов на основе RGB- и 3D-модели.

В работе [51] представлена гибридная архитектура CNN для распознавания жестов в формате RGB-D, нейросетевая модель, которая извлекает пространственно-временные характеристики двух модальностей с использованием комбинации 3D CNN и LSTM. Временная информация входных данных модели кодируется с использованием двухпоточной архитектуры, основанной на уменьшенной модели сети ResNet-10.

Выводы по разделу

Как показано, основные направления интеллектуального анализа жестов и жестовой речи охватывают отслеживание и определение конфигурации рук, анализ временных и пространственных характеристик жеста, а также цепочек жестов (рис. 2).

Методы оценки конфигурации руки можно разделить на 2D и 3D. Большинство современных моделей ориентированы на работу с 3D-данными; тезис о том, что карты глубины больше подходят для решения этой задачи, выдвигается и в ряде обзоров, например [1, 41]. В то же время очевидно, что возможности CNN ограничиваются работой со статическими изображениями, и эту архитектуру необходимо объединить с LSTM для решения задачи оценки конфигурации в динамических сценах. Эффективность CNN существенно повышается в том случае, если в процессе обучения модели используются предварительные знания о геометрии руки [52].

■ ***Рис. 2.*** Основные задачи из области автоматического распознавания жестов
■ ***Fig. 2.*** The main tasks in the field of automatic gesture recognition

Наилучшие результаты для слитной речи показывают модели, основанные на архитектурах типа CNN + LSTM. Основной проблемой распознавания слитной речи является отсутствие предварительных знаний о границах жестов, т. е. временной разметки. Как следует из настоящего обзора, именно комбинация CNN и LSTM позволяет решить последнюю проблему. Из табл. 3 видно, что наиболее перспективным оказывается совмещение модальностей.

Заключение

В настоящей статье были рассмотрены современные методы и подходы к задаче распознавания жестов и жестовых языков. Можно сформулировать ряд проблем, решить которые необходимо для создания системы распознавания жестовых языков.

1. Одной из основных модальностей жестовых языков является визуальная модальность. Из этого следует, что эффективная система распознавания жестового языка требует решения ряда задач из области компьютерного зрения. К основным проблемам из этой области относятся шумы,

конволюции, вариации размеров и оттенков артикуляторов, динамическая освещенность.

2. В настоящий момент основные задачи, стоящие перед создателями систем автоматического распознавания жестов, сводятся к поиску эффективных методов детектирования и распознавания конфигурации рук, определения конфигурации артикуляторов, распознавания динамических жестов в потоке речи.

В последние годы был создан ряд методов, позволяющих достаточно уверенно решать первые две проблемы. Анализ применения глубокого машинного обучения для решения третьей задачи позволяет сформулировать следующую проблему.

3. Немногочисленность корпусов жестовых языков в открытом доступе. Сбор корпусов жестовых языков достаточно трудоемок из-за относительно небольшого количества записей и отсутствия общепринятых систем нотации жестовых языков.

На основании представленного обзора авторы делают вывод, что наиболее эффективным способом решения проблем, обозначенных в п. 1, будет применение моделей, работающих с данными в формате 3D. Использование массива сенсоров, поддерживающих создание карты глубины, позволяет точно моделировать кисть руки в трехмерном пространстве, а наличие больших баз данных дает возможность проводить автоматическую разметку полученного набора данных.

Анализ современных подходов позволяет с достаточной уверенностью утверждать, что эффективным решением задач, обозначенных в п. 2, будет одновременное извлечение пространственных и временных признаков жеста с использованием комбинации архитектур CNN и LSTM. 3D сверточная LSTM-сеть за счет хранения 3D пространственной информации может формировать более эффективные пространственно-временные характеристики жеста.

Наконец, решение проблемы набора данных, в конечном счете, сильно зависит от жестового языка, с которым работает исследователь. Авторы настоящей статьи представили собственный набор данных для русского жестового языка в трехмерном формате [53]; впоследствии этот набор данных использовался для системы автоматического распознавания жестов на русском жестовом языке [54–56].

Финансовая поддержка

Аналитический обзор методов и решений, применяемых для распознавания жестов при помощи глубокого обучения, выполнен за счет гранта Российского научного фонда (№ 21-71-00141, https://rscf.ru/project/21-71-00141/), исследование жестовых корпусов и наборов данных проведено в рамках бюджетной темы № 0073-2019-0005.

Литература

1. **Koller O.** Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv: 2008.09918*, 2020.
2. **Rastgoo R., Kiani K., Escalera S.** Sign language recognition: A deep survey. *Expert Systems with Applications*, 2021, vol. 164, 113794. doi:10.1016/j.eswa.2020.113794
3. **Koller O., Forster J., Ney H.** Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 2015, vol. 141, pp. 108–125. doi:10.1016/j.cviu.2015.09.013
4. **Wan J., Li S. Z., Zhao Y., Zhou S., Guyon I., Escalera S.** ChaLearn looking at people RGB-D isolated and continuous datasets for gesture recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 761–769. doi:10.1109/CVPRW.2016.100
5. **Madadi M., Escalera S., Carruesco A., Andujar C., Baró X., Gonzàlez J.** Occlusion aware hand pose recovery from sequences of depth images. *Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017, pp. 230–237. doi:10.1109/FG.2017.37
6. **Yuan S., Ye Q., Stenger B., Jain S., Kim. T.** BigHand2.2M benchmark: Hand pose dataset and state of the art analysis. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2605–2613. doi:10.1109/CVPR.2017.279
7. **Zhang Y., Cao C., Cheng J., Lu H.** EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition. *Proceedings of the IEEE Transactions on Multimedia (T-MM)*, 2018, vol. 20, no. 5, pp. 1038–1050. doi:10.1109/TMM.2018.2808769
8. **Camgoz N., Hadfield S., Koller S., Ney H., Bowden R.** Neural sign language translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7784–7793. doi:10.1109/CVPR.2018.00812
9. **Rastgoo R., Kiani K., Escalera S.** Hand sign language recognition using multi-view hand skeleton. *Expert Systems with Applications*, 2020, vol. 150, ID 113336. doi:10.1016/j.eswa.2020.113336
10. **Moon G., Yu S. I., Wen H., Shiratori T., Lee K. M.** InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image.

Proceedings of the Computer Vision — ECCV 2020: 16th European Conference, Glasgow, 2020, pp. 548–564. doi:10.1007/978-3-030-58565-5_33

11. **Li D., Rodriguez Ch., Yu X.** Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 1459–1469. doi:10.1109/WACV45572.2020.9093512
12. **Sincan O. M., Keles H. Y.** AUTSL: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 2020 vol. 8, pp. 181340–181355. doi:10.1109/ACCESS.2020.3028072
13. **Starmans M., Voort van der S., Castillo Tovar J., Veenland J., Klein S., Niessen W.** Radiomics: Data Mining Using Quantitative Medical Image Features. In: *Handbook of Medical Image Computing and Computer Assisted Intervention*. Kevin Zhou S., Daniel Rueckert D., Fichtinger G. (Eds.). Elsevier Science, 2020. Pp. 429–456. doi:10.1016/B978-0-12-816176-0.00023-5
14. **Perimal M., Basah S. N., Safar M. J. A., Yazid H.** Hand-gesture recognition-algorithm based on finger counting. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 2018, vol. 10(1–13), pp. 19–24.
15. **Ren S., He K., Girshick R., Sun J.** Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, vol. 39(6), pp. 1137–1149. doi:10.1109/TPAMI.2016.2577031
16. **Li W.** Analysis of object detection performance based on Faster R-CNN. *Journal of Physics: Conference Series*, 2021, vol. 1827, ID 012085. doi:10.1088/1742-6596/1827/1/012085
17. **Redmon J., Divvala S., Girshick R., Farhadi A.** You Only Look Once: unified, real-time object detection. *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. doi:10.1109/CVPR.2016.91
18. **Liu W., Anguelov D., Erhan D., Szegedy Ch., Reed S., Fu Ch.-Y., Berg A. C.** SSD: Single Shot MultiBox Detector. In: *Computer Vision — ECCV 2016*. Leibe B., Matas J., Sebe N., Welling M. (Eds). Lecture Notes in Computer Science. Springer, Cham, 2016. Vol. 9905. Pp. 21–37. doi:10.1007/978-3-319-46448-0_2
19. **Haque A., Peng B., Luo Z., Alahi A., Yeung S., Fei-Fei L.** Towards Viewpoint Invariant 3D Human Pose Estimation. In: *Computer Vision — ECCV 2016*. Leibe B., Matas J., Sebe N., Welling M. (Eds). Lecture Notes in Computer Science. Springer, 2016. Vol. 9905. Pp. 160–177. doi:10.1007/978-3-319-46448-0_10
20. **Marín-Jiménez M., Romero-Ramirez F., Muñoz-Salinas R., Medina-Carnicer R.** 3D human pose estimation from depth maps using a deep combination of poses. *Journal of Visual Communication and Image Representation*, 2018, vol. 55, pp. 627–639. doi:10.1016/j.jvcir.2018.07.010
21. **Zhou Y., Jiang G., Lin Y.** A novel finger and hand pose estimation technique for real-time hand gesture recognition. *Pattern Recognition*, 2016, vol. 49, pp. 102–114. doi:10.1016/j.patcog.2015.07.014
22. **Molina J., Pajuelo J. A., Martínez J. M.** Real-time motion-based hand gestures recognition from time-of-flight video. *Journal of Signal Processing Systems*, 2017, vol. 86, no. 1, pp. 17–25. doi:10.1007/s11265-015-1090-5
23. **Devineau G., Moutarde F., Xi W., Yang J.** Deep learning for hand gesture recognition on skeletal data. *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an, China, 2018, pp. 106–113. doi:10.1109/FG.2018.00025
24. **Ma X., Peng J.** Kinect sensor-based long-distance hand gesture recognition and fingertip detection with depth information. *Journal of Sensors*, 2018, vol. 2018, ID 5809769. doi:10.1155/2018/5809769
25. **Tekin B., Bogo F., Pollefeys M.** H+O: Unified egocentric recognition of 3D hand-object poses and interactions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4511–4520. doi:10.1109/CVPR.2019.00464
26. **Raheja J. L., Mishra A., Chaudhary A.** Indian sign language recognition using SVM. *Pattern Recognition and Image Analysis*, 2016, vol. 26, no. 2, pp. 434–441. doi:10.1134/S1054661816020164
27. **Grif M. G., Kugaevskikh A. V.** Recognition of deaf gestures based on a bio-inspired neural network. *Journal of Physics: Conference Series*, 2020, vol. 1661, ID 012038. doi:10.1088/1742-6596/1661/1/012038
28. **Dibra E., Melchior S., Balkis A., Wolf T., Oztireli C., Gross M.** Monocular RGB hand pose inference from unsupervised refinable nets. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1075–1085. doi:10.1109/CVPRW.2018.00155
29. **Sinha A., Choi C., Ramani K.** Deephand: Robust hand pose estimation by completing a matrix imputed with deep features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4150–4158. doi:10.1109/CVPR.2016.450
30. **Wei S., Ramakrishna V., Kanade T., Sheikh Y.** Convolutional pose machines. *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4724–4732. doi:10.1109/CVPR.2016.511
31. **Zimmermann C., Brox T.** Learning to estimate 3D hand pose from single RGB images. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4903–4911. doi:10.1109/ICCV.2017.525
32. **Mueller F., Bernard F., Sotnychenko O., Mehta D., Sridhar S., Casas D., Theobalt C.** GANerated hands for real-time 3D hand tracking from monocular RGB. *Proceedings of the IEEE Conference on Computer Vi-*

sion and Pattern Recognition (CVPR), 2018, pp. 49–59. doi:10.1109/CVPR.2018.00013

33. **Baek S., Kim K. I., Kim T. K.** Augmented skeleton space transfer for depth-based hand pose estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8330–8339. doi:10.1109/CVPR.2018.00869
34. **Zhu J. Y., Park T., Isola P., Efros A. A.** Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2223–2232. doi:10.1109/ICCV.2017.244
35. **Ge L., Liang H., Yuan J., Thalmann D.** 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1991–2000. doi:10.1109/CVPR.2017.602
36. **Tang D., Chang H. J., Tejani A., Kim T.** Latent regression forest: structured estimation of 3D articulated hand posture. *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3786–3793. doi:10.1109/CVPR.2014.490
37. **Spurr A., Song J., Park S., Hilliges O.** Cross-modal deep variational hand pose estimation. *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 89–98. doi:10.1109/CVPR.2018.00017
38. **Moon G., Chang J., Lee K.** V2V-PoseNet: Voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5079–5088. doi:10.1109/CVPR.2018.00533
39. **Li Y., Xue Z., Wang Y., Ge L., Ren Z., Rodriguez J.** End-to-end 3D hand pose estimation from stereo cameras. *Proceedings of the British Machine Vision Conference (BMVC-2019)*, 2019, pp. 38.11–38.13. doi:10.5244/C.33.38
40. **Gomez-Donoso F., Orts-Escolano S., Cazorla M.** Accurate and efficient 3D hand pose regression for robot hand teleoperation using a monocular RGB camera. *Expert Systems with Applications*, 2019, vol. 136, pp. 327–337. doi:10.1016/j.eswa.2019.06.055
41. **Supancic J., Rogez G., Yang Y., Shotton J., Ramana D.** Depth-based hand pose estimation: methods, data, and challenges. *International Journal of Computer Vision*, 2018, vol. 126, no. 11, pp. 1180–1198. doi:10.1007/s11263-018-1081-7
42. **Sun X., Wei Y., Liang S., Tang X., Sun J.** Cascaded hand pose regression. *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 824–832. doi:10.1109/CVPR.2015.7298683
43. **Zhang J., Jiao J., Chen M., Qu L., Xu X., Yang Q.** A hand pose tracking benchmark from stereo matching. *Proceedings of 2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 982–986. doi:10.1109/ICIP.2017.8296428
44. **Choudhury A., Talukdar A. K., Sarma K. K., Bhuyan M. K.** An adaptive thresholding-based movement epenthesis detection technique using hybrid feature set for continuous fingerspelling recognition. *SN Computer Science*, 2021, vol. 2(128). doi:10.1007/s42979-021-00544-5
45. **Koller O., Camgoz N. C., Ney H., Bowden R.** Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, vol. 42, no. 9, pp. 2306–2320. doi:10.1109/TPAMI.2019.2911077
46. **Huang J., Zhou W., Zhang Q., Li H., Li W.** Video-based sign language recognition without temporal segmentation. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 2257–2264.
47. **Joze H. R. V., Koller O.** MS-ASL: A large-scale data set and benchmark for understanding American sign language. *arXiv preprint arXiv:1812.01053*, 2018.
48. **Cui R., Liu H., Zhang C.** A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 2019, vol. 21, no. 7, pp. 1880–1891. doi:10.1109/TMM.2018.2889563
49. **Wei C., Zhou W., Pu J., Li H.** Deep grammatical multi-classifier for continuous sign language recognition. *Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 2019, pp. 435–442. doi:10.1109/BigMM.2019.00027
50. **Elboushaki A., Hannane R., Afdel K., Koutti L.** MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences. *Expert Systems with Applications*, 2020, vol. 139, ID 112829. doi:10.1016/j.eswa.2019.112829
51. **Xu C., Zhou J., Cai W., Jiang Y., Li Y., Liu Y.** Robust 3D hand detection from a single RGB-D image in unconstrained environments. *Sensors*, 2020, vol. 20, no. 21, ID 6360. doi:10.3390/s20216360
52. **Zhou X., Wan Q., Zhang W., Xue X., Wei Y.** Model-based deep hand pose estimation. *arXiv preprint arXiv:1606.06854*, 2016.
53. **Кагиров И. А., Рюмин Д. А., Аксёнов А. А., Карпов А. А.** Мультимедийная база данных жестов русского жестового языка в трехмерном формате. *Вопросы языкознания*, 2020, № 1, с. 104–123. doi:10.31857/S0373658X0008302-1
54. **Kagirov I., Ryumin D., Axyonov A.** Method for multimodal recognition of one-handed sign language gestures through 3D convolution and LSTM neural networks. *Proceedings of the 21th International Conference on Speech and Computer (SPECOM-2019)*, Springer, LNAI, 2019, vol. 116582019, pp. 191–200. doi:10.1007/978-3-030-26061-3_20

55. **Рюмин Д.** Метод автоматического видеоанализа движений рук и распознавания жестов в человеко-машинных интерфейсах. *Научно-технический вестник информационных технологий, механики и оптики*, 2020, т. 20, № 4, с. 525–531. doi:10.17586/2226-1494-2020-20-4-525-531

56. **Ryumin D., Kagirov I., Axyonov A., Pavlyuk N., Saveliev A., Kipyatkova I., Zelezny M., Mporas I., Karpov A.** A Multimodal user interface for an assistive robotic shopping cart. *Electronics*, 2020, vol. 9, no. 12, ID 2093, pp. 1–25. doi:10.3390/electronics9122093

UDC 004.93′12
doi:10.31799/1684-8853-2021-6-10-20

Analytical review of models and methods for automatic recognition of gestures and sign languages

D. A. Ryumin[a], PhD, Tech., Senior Research, orcid.org/0000-0002-7935-0569, ryumin.d@iias.spb.su
I. A. Kagirov[a], Research Fellow, orcid.org/0000-0003-1196-1117, kagirov@iias.spb.su
A. A. Axyonov[a], Junior Research, orcid.org/0000-0002-7479-2851
A. A. Karpov[a], Dr. Sc., Tech., Associate Professor, Principal Researcher, orcid.org/0000-0003-3424-652X
[a]St. Petersburg Federal Research Center of the RAS, 39, 14th Line, 199178, Saint-Petersburg, Russian Federation

Introduction: Currently, the recognition of gestures and sign languages is one of the most intensively developing areas in computer vision and applied linguistics. The results of current investigations are applied in a wide range of areas, from sign language translation to gesture-based interfaces. In that regard, various systems and methods for the analysis of gestural data are being developed. **Purpose:** A detailed review of methods and a comparative analysis of current approaches in automatic recognition of gestures and sign languages. **Results:** The main gesture recognition problems are the following: detection of articulators (mainly hands), pose estimation and segmentation of gestures in the flow of speech. The authors conclude that the use of two-stream convolutional and recurrent neural network architectures is generally promising for efficient extraction and processing of spatial and temporal features, thus solving the problem of dynamic gestures and coarticulations. This solution, however, heavily depends on the quality and availability of data sets. **Practical relevance:** This review can be considered a contribution to the study of rapidly developing sign language recognition, irrespective to particular natural sign languages. The results of the work can be used in the development of software systems for automatic gesture and sign language recognition.

Keywords — sign languages, gesture corpus, gesture recognition, hand geometry, machine learning, computer vision, neural network model, digital processing, morphological processing, informative features.

For citation: Ryumin D. A., Kagirov I. A., Axyonov A. A., Karpov A. A. Analytical review of models and methods for automatic recognition of gestures and sign languages. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2021, no. 6, pp. 10–20 (In Russian). doi:10.31799/1684-8853-2021-6-10-20

Financial support

The analytical review of methods and solutions applied for deep learning-based gesture recognition was funded by Russian Science Foundation (No 21-71-00141, https://rscf.ru/project/21-71-00141), the survey of gesture corpora and datasets was carried out within the framework of the budgetary theme No 0073-2019-0005.

References

1. Koller O. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020.
2. Rastgoo R., Kiani K., Escalera S. Sign language recognition: A deep survey. *Expert Systems with Applications*, 2021, vol. 164, 113794. doi:10.1016/j.eswa.2020.113794
3. Koller O., Forster J., Ney H. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 2015, vol. 141, pp. 108–125. doi:10.1016/j.cviu.2015.09.013
4. Wan J., Li S. Z., Zhao Y., Zhou S., Guyon I., Escalera S. ChaLearn looking at people RGB-D isolated and continuous datasets for gesture recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 761–769. doi:10.1109/CVPRW.2016.100
5. Madadi M., Escalera S., Carruesco A., Andujar C., Baró X., Gonzàlez J. Occlusion aware hand pose recovery from sequences of depth images. *Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017, pp. 230–237. doi:10.1109/FG.2017.37
6. Yuan S., Ye Q., Stenger B., Jain S., Kim. T. BigHand2.2M benchmark: Hand pose dataset and state of the art analysis. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2605–2613. doi:10.1109/CVPR.2017.279
7. Zhang Y., Cao C., Cheng J., Lu H. EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition. *Proceedings of the IEEE Transactions on Multimedia (T-MM)*, 2018, vol. 20, no. 5, pp. 1038–1050. doi:10.1109/TMM.2018.2808769
8. Camgoz N., Hadfield S., Koller S., Ney H., Bowden R. Neural sign language translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7784–7793. doi:10.1109/CVPR.2018.00812
9. Rastgoo R., Kiani K., Escalera S. Hand sign language recognition using multi-view hand skeleton. *Expert Systems with Applications*, 2020, vol. 150, ID 113336. doi:10.1016/j.eswa.2020.113336
10. Moon G., Yu S. I., Wen H., Shiratori T., Lee K. M. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. *Proceedings of the Computer Vision — ECCV 2020: 16th European Conference*, Glasgow, 2020, pp. 548–564. doi:10.1007/978-3-030-58565-5_33
11. Li D., Rodriguez Ch., Yu X. Word-level deep sign language recognition from video: A new large-scale dataset and meth-

ods comparison. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 1459–1469. doi:10.1109/WACV45572.2020.9093512
12. Sincan O. M., Keles H. Y. AUTSL: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 2020 vol. 8, pp. 181340–181355. doi:10.1109/ACCESS.2020.3028072
13. Starmans M., Voort van der S., Castillo Tovar J., Veenland J., Klein S., Niessen W. *Radiomics: Data Mining Using Quantitative Medical Image Features*. In: *Handbook of Medical Image Computing and Computer Assisted Intervention*. Kevin Zhou S., Daniel Rueckert D., Fichtinger G. (Eds.). Elsevier Science, 2020. Pp. 429–456. doi:10.1016/B978-0-12-816176-0.00023-5
14. Perimal M., Basah S. N., Safar M. J. A., Yazid H. Hand-gesture recognition-algorithm based on finger counting. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 2018, vol. 10(1-13), pp. 19–24.
15. Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, vol. 39(6), pp. 1137–1149. doi:10.1109/TPAMI.2016.2577031
16. Li W. Analysis of object detection performance based on Faster R-CNN. *Journal of Physics: Conference Series*, 2021, vol. 1827, ID 012085. doi:10.1088/1742-6596/1827/1/012085
17. Redmon J., Divvala S., Girshick R., Farhadi A. You Only Look Once: unified, real-time object detection. *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788. doi:10.1109/CVPR.2016.91
18. Liu W., Anguelov D., Erhan D., Szegedy Ch., Reed S., Fu Ch.-Y., Berg A. C. *SSD: Single Shot MultiBox Detector*. In: *Computer Vision — ECCV 2016*. Leibe B., Matas J., Sebe N., Welling M. (Eds). Lecture Notes in Computer Science. Springer, Cham, 2016. Vol. 9905. Pp. 21–37. doi:10.1007/978-3-319-46448-0_2
19. Haque A., Peng B., Luo Z., Alahi A., Yeung S., Fei-Fei L. *Towards Viewpoint Invariant 3D Human Pose Estimation*. In: *Computer Vision — ECCV 2016*. Leibe B., Matas J., Sebe N., Welling M. (Eds). Lecture Notes in Computer Science. Springer, 2016. Vol. 9905. Pp. 160–177. doi:10.1007/978-3-319-46448-0_10
20. Marín-Jiménez M., Romero-Ramirez F., Muñoz-Salinas R., Medina-Carnicer R. 3D human pose estimation from depth maps using a deep combination of poses. *Journal of Visual Communication and Image Representation*, 2018, vol. 55, pp. 627–639. doi:10.1016/j.jvcir.2018.07.010
21. Zhou Y., Jiang G., Lin Y. A novel finger and hand pose estimation technique for real-time hand gesture recognition. *Pattern Recognition*, 2016, vol. 49, pp. 102–114. doi:10.1016/j.patcog.2015.07.014
22. Molina J., Pajuelo J. A., Martínez J. M. Real-time motion-based hand gestures recognition from time-of-flight video. *Journal of Signal Processing Systems*, 2017, vol. 86, no. 1, pp. 17–25. doi:10.1007/s11265-015-1090-5
23. Devineau G., Moutarde F., Xi W., Yang J. Deep learning for hand gesture recognition on skeletal data. *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an, China, 2018, pp. 106–113. doi:10.1109/FG.2018.00025
24. Ma X., Peng J. Kinect sensor-based long-distance hand gesture recognition and fingertip detection with depth information. *Journal of Sensors*, 2018, vol. 2018, ID 5809769. doi:10.1155/2018/5809769
25. Tekin B., Bogo F., Pollefeys M. H+O: Unified egocentric recognition of 3D hand-object poses and interactions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4511–4520. doi:10.1109/CVPR.2019.00464
26. Raheja J. L., Mishra A., Chaudhary A. Indian sign language recognition using SVM. *Pattern Recognition and Image Analysis*, 2016, vol. 26, no. 2, pp. 434–441. doi:10.1134/S1054661816020164
27. Grif M. G., Kugaevskikh A. V. Recognition of deaf gestures based on a bio-inspired neural network. *Journal of Physics: Conference Series*, 2020, vol. 1661, ID 012038. doi:10.1088/1742-6596/1661/1/012038
28. Dibra E., Melchior S., Balkis A., Wolf T., Oztireli C., Gross M. Monocular RGB hand pose inference from unsupervised refinable nets. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1075–1085. doi:10.1109/CVPRW.2018.00155
29. Sinha A., Choi C., Ramani K. Deephand: Robust hand pose estimation by completing a matrix imputed with deep features. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4150–4158. doi:10.1109/CVPR.2016.450
30. Wei S., Ramakrishna V., Kanade T., Sheikh Y. Convolutional pose machines. *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4724–4732. doi: 10.1109/CVPR.2016.511
31. Zimmermann C., Brox T. Learning to estimate 3D hand pose from single RGB images. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4903–4911. doi:10.1109/ICCV.2017.525
32. Mueller F., Bernard F., Sotnychenko O., Mehta D., Sridhar S., Casas D., Theobalt C. GANerated hands for real-time 3D hand tracking from monocular RGB. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 49–59. doi:10.1109/CVPR.2018.00013
33. Baek S., Kim K. I., Kim T. K. Augmented skeleton space transfer for depth-based hand pose estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8330–8339. doi:10.1109/CVPR.2018.00869
34. Zhu J. Y., Park T., Isola P., Efros A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2223–2232. doi:10.1109/ICCV.2017.244
35. Ge L., Liang H., Yuan J., Thalmann D. 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1991–2000. doi:10.1109/CVPR.2017.602
36. Tang D., Chang H. J., Tejani A., Kim T. Latent regression forest: structured estimation of 3D articulated hand posture. *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3786–3793. doi:10.1109/CVPR.2014.490
37. Spurr A., Song J., Park S., Hilliges O. Cross-modal deep variational hand pose estimation. *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 89–98. doi:10.1109/CVPR.2018.00017
38. Moon G., Chang J., Lee K. V2V-PoseNet: Voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5079–5088. doi:10.1109/CVPR.2018.00533
39. Li Y., Xue Z., Wang Y., Ge L., Ren Z., Rodriguez J. End-to-end 3D hand pose estimation from stereo cameras. *Proceedings of the British Machine Vision Conference (BMVC-2019)*, 2019, pp. 38.11–38.13. doi:10.5244/C.33.38
40. Gomez-Donoso F., Orts-Escolano S., Cazorla M. Accurate and efficient 3D hand pose regression for robot hand teleoperation using a monocular RGB camera. *Expert Systems with Applications*, 2019, vol. 136, pp. 327–337. doi:10.1016/j.eswa.2019.06.055
41. Supancic J., Rogez G., Yang Y., Shotton J., Ramana D. Depth-based hand pose estimation: methods, data, and challenges. *International Journal of Computer Vision*, 2018, vol. 126, no. 11, pp. 1180–1198. doi:10.1007/s11263-018-1081-7
42. Sun X., Wei Y., Liang S., Tang X., Sun J. Cascaded hand pose regression. *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 824–832. doi:10.1109/CVPR.2015.7298683
43. Zhang J., Jiao J., Chen M., Qu L., Xu X., Yang Q. A hand pose tracking benchmark from stereo matching. *Proceedings of 2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 982–986. doi:10.1109/ICIP.2017.8296428
44. Choudhury A., Talukdar A. K., Sarma K. K., Bhuyan M. K. An adaptive thresholding-based movement epenthesis detection technique using hybrid feature set for continuous fingerspelling recognition. *SN Computer Science*, 2021, vol. 2(128). doi:10.1007/s42979-021-00544-5
45. Koller O., Camgoz N. C., Ney H., Bowden R. Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, vol. 42, no. 9, pp. 2306–2320. doi:10.1109/TPAMI.2019.2911077
46. Huang J., Zhou W., Zhang Q., Li H., Li W. Video-based sign language recognition without temporal segmentation. *Pro-*

ceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 2257–2264.
47. Joze H. R. V., Koller O. MS-ASL: A large-scale data set and benchmark for understanding American sign language. *arXiv preprint arXiv:1812.01053*, 2018.
48. Cui R., Liu H., Zhang C. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 2019, vol. 21, no. 7, pp. 1880–1891. doi:10.1109/TMM.2018.2889563
49. Wei C., Zhou W., Pu J., Li H. Deep grammatical multi-classifier for continuous sign language recognition. *Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 2019, pp. 435–442. doi:10.1109/BigMM.2019.00027
50. Elboushaki A., Hannane R., Afdel K., Koutti L. MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences. *Expert Systems with Applications*, 2020, vol. 139, ID 112829. doi:10.1016/j.eswa.2019.112829
51. Xu C., Zhou J., Cai W., Jiang Y., Li Y., Liu Y. Robust 3D hand detection from a single RGB-D image in unconstrained environments. *Sensors*, 2020, vol. 20, no. 21, ID 6360. doi:10.3390/s20216360
52. Zhou X., Wan Q., Zhang W., Xue X., Wei Y. Model-based deep hand pose estimation. *arXiv preprint arXiv:1606.06854*, 2016.
53. Kagirov I. A., Ryumin D. A., Axyonov A. A., Karpov A. A. Multimedia database of Russian sign language items in 3D. *Voprosy jazykoznanija*, 2020, vol. 1, pp. 104–123 (In Russian). doi:10.31857/S0373658X0008302-1
54. Kagirov I., Ryumin D., Axyonov A. Method for multimodal recognition of one-handed sign language gestures through 3D convolution and LSTM neural networks. *Proceedings of the 21th International Conference on Speech and Computer (SPECOM-2019)*, Springer, LNAI, 2019, vol. 116582019, pp. 191–200. doi:10.1007/978-3-030-26061-3_20
55. Ryumin D. Automated hand detection method for tasks of gesture recognition in human-machine interfaces. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2020, vol. 20, no. 4, pp. 525–531 (In Russian). doi:10.17586/2226-1494-2020-20-4-525-531
56. Ryumin D., Kagirov I., Axyonov A., Pavlyuk N., Saveliev A., Kipyatkova I., Zelezny M., Mporas I., Karpov A. A multimodal user interface for an assistive robotic shopping cart. *Electronics*, 2020, vol. 9, no. 12, ID 2093, pp. 1–25. doi:10.3390/electronics9122093