

Multimodal Data Analysis in Emotion Recognition: A Review

D. A. Berdyshev¹, A. G. Shishkin¹DOI: [10.18255/1818-1015-2025-3-252-281](https://doi.org/10.18255/1818-1015-2025-3-252-281)¹Lomonosov Moscow State University, Moscow, Russia

MSC2020: 68T07

Research article

Full text in Russian

Received May 5, 2025

Revised June 6, 2025

Accepted June 11, 2025

The use of multimodal data in emotion recognition systems has great potential for applications in various fields: health-care, human-machine interfaces, operator monitoring, and marketing. Until recently, the development of emotion recognition systems based on multimodal data was constrained by insufficient computing power. However, with the advent of high-performance GPU-based systems and the development of efficient deep neural network architectures, there has been a surge of research aimed at using multiple modalities such as audio, video, and physiological signals to accurately detect human emotions. In addition, physiological data from wearable devices has become important due to the relative ease of its collection and the accuracy it enables. This paper discusses architectures and methods for applying deep neural networks to analyse multimodal data to improve the accuracy and reliability of emotion recognition systems, presenting current approaches to implementing such algorithms and existing open multimodal datasets.

Keywords: emotion recognition; multimodal data; neural networks; machine learning

INFORMATION ABOUT THE AUTHORS

Berdyshev, Daniil A. (corresponding author)	ORCID iD: 0009-0006-4484-3853 . E-mail: danberdyshev@gmail.com Post-graduate Student
Shishkin, Aleksei G.	ORCID iD: 0000-0002-6170-8630 . E-mail: shishkin@cs.msu.ru Dr. Sc., Professor

For citation: D. A. Berdyshev and A. G. Shishkin, "Multimodal data analysis in emotion recognition: a review", *Modeling and Analysis of Information Systems*, vol. 32, no. 3, pp. 252–281, 2025. DOI: [10.18255/1818-1015-2025-3-252-281](https://doi.org/10.18255/1818-1015-2025-3-252-281).

Анализ мультимодальных данных в распознавании эмоций

Д. А. Бердышев¹, А. Г. Шишкин¹

DOI: [10.18255/1818-1015-2025-3-252-281](https://doi.org/10.18255/1818-1015-2025-3-252-281)

¹Московский государственный университет им. М.В. Ломоносова, Москва, Россия

УДК 004.93

Научная статья

Полный текст на русском языке

Получена 5 мая 2025 г.

После доработки 6 июня 2025 г.

Принята к публикации 11 июня 2025 г.

Использование мультимодальных данных в системах распознавания эмоций имеет огромный потенциал для приложений в различных областях: здравоохранение, человеко-машинные интерфейсы, контроль состояния операторов, маркетинг. До недавнего времени развитие систем распознавания эмоций на основе мультимодальных данных сдерживалось недостаточной мощностью вычислительной техники. Однако с появлением высокопроизводительных систем на основе графических процессоров и разработкой эффективных архитектур глубоких нейронных сетей произошел всплеск исследований, направленных на использование нескольких модальностей, таких как аудио, видео и физиологические сигналы, для точного определения человеческих эмоций. Помимо этого, немаловажную роль стали играть физиологические данные, полученные с помощью носимых устройств, благодаря относительной простоте их сбора и точности, которую они позволяют достигать. В данной статье рассмотрены архитектуры и методы применения глубоких нейронных сетей для анализа мультимодальных данных с целью повышения точности и надежности систем распознавания эмоций, представлены современные подходы к реализации таких алгоритмов и существующие открытые наборы мультимодальных данных.

Ключевые слова: распознавание эмоций; мультимодальные данные; нейронные сети; машинное обучение

ИНФОРМАЦИЯ ОБ АВТОРАХ

Бердышев, Даниил Алексеевич
(автор для корреспонденции)

ORCID iD: [0009-0006-4484-3853](https://orcid.org/0009-0006-4484-3853). E-mail: danberdyshev@gmail.com
Аспирант

Шишкин, Алексей Геннадиевич

ORCID iD: [0000-0002-6170-8630](https://orcid.org/0000-0002-6170-8630). E-mail: shishkin@cs.msu.ru
Доктор физ.-мат. наук, профессор

Для цитирования: D. A. Berdyshev and A. G. Shishkin, "Multimodal data analysis in emotion recognition: a review", *Modeling and Analysis of Information Systems*, vol. 32, no. 3, pp. 252–281, 2025. DOI: [10.18255/1818-1015-2025-3-252-281](https://doi.org/10.18255/1818-1015-2025-3-252-281).

Введение

Распознавание эмоций — важный аспект человеческого общения и взаимодействия. Традиционно, системы распознавания эмоций основывались на одномодальных данных, таких как выражения лица [1] или речевые сигналы [2]. Однако прогресс в развитии вычислительной техники, позволивший использовать большое число мощных графических ускорителей, открыл возможности интеграции множества модальностей, что привело к появлению мультимодальных систем распознавания эмоций [3]. В добавление к этому, применение глубоких нейронных сетей произвело революцию в этой области, предложив мощные инструменты для анализа мультимодальных данных с целью распознавания человеческих эмоций [4].

Одним из основных преимуществ мультимодальных данных является их способность более полно и детально отражать человеческие эмоции [5]. Комбинируя информацию из разных модальностей, таких как выражения лица и язык тела, речевые и физиологические сигналы, мультимодальные системы распознавания эмоций способны преодолеть ограничения классических одномодальных подходов, которые обычно очень чувствительны к шуму и неоднозначности данных. Например, выражения лица, которые, в целом, представляют ценную информацию об эмоциональном состоянии человека, не всегда могут быть надёжными источниками данных в ситуациях, когда люди маскируют свои эмоции, или демонстрируют тонкие невербальные сигналы. В подобных случаях интеграция также и других модальностей, таких как тон голоса или физиологические реакции, может предоставить дополнительную информацию, которая позволит повысить общую точность и надёжность системы распознавания эмоций.

В последние годы глубокие нейронные сети (DNN) стали мощным инструментом для анализа мультимодальных данных, позволяя извлекать сложные закономерности и отношения в различных модальностях [6]. Сверточные нейронные сети (CNN) [7] и рекуррентные нейронные сети (RNN) [8] являются широко используемыми архитектурами для задач мультимодального распознавания эмоций. CNN хорошо подходят для обработки пространственных данных, таких как изображения или выражения лица, а RNN успешно применяются для обработки временных данных, таких как речевые сигналы или физиологические данные. Объединив эти архитектуры в мультимодальную структуру, исследователи могут использовать сильные стороны каждой модальности для улучшения эффективности распознавания эмоций. Один из способов такого объединения — нейронные сети с архитектурой «трансформер» [9], которая позволяет использовать свой кодировщик для каждой из модальностей.

Несмотря на многообещающие перспективы мультимодального распознавания эмоций, в этой области существует целый ряд проблем [10]. Одной из основных является объединение информации из отдельных модальностей, поскольку каждая модальность может иметь разную временную динамику и уровни шума. Поэтому активной областью исследований является разработка эффективных стратегий слияния, которые могут интегрировать информацию из различных модальностей, сохраняя при этом соответствующие им особенности [10, 11]. Помимо этого, сбор и аннотирование мультимодальных наборов данных сложны и дорогостоящи, поскольку зачастую требуют специального оборудования и опыта [12].

Однако вместе с этими проблемами появляются возможности для инноваций и развития. Поскольку современные технологии продолжают непрерывно развиваться, доступность носимых устройств и различных датчиков позволяет собирать мультимодальные данные в реальных условиях [13], что дает возможность исследователям изучать эмоции в естественной среде. Более того, недавние разработки в области обучения с переносом (Transfer Learning) и обучения без учителя открывают возможности для использования предварительно обученных моделей и методов обучения без учителя для улучшения обобщающей способности и масштабируемости мультимодальных систем распознавания эмоций [14].

В настоящей работе выполнен анализ современного состояния мультимодальных систем распознавания эмоций и соответствующих наборов данных. В разделе 1 рассмотрены доступные на данный момент мультимодальные наборы данных эмоциональных состояний. В разделе 2 рассмотрены современные подходы классификации эмоциональных состояний.

1. Данные

Обучение нейронных сетей для решения прикладных задач, в том числе, мультимодального распознавания эмоций, требует огромного количества размеченных данных [15]. Из-за сложности и разнообразия эмоциональных выражений в разных модальностях получение таких данных сопряжено с рядом серьезных проблем. Во-первых, сбор данных по множеству модальностей, таких как выражение лица, речевые сигналы, физиологические реакции и др., требует обширных ресурсов, времени и усилий. Во-вторых, аннотирование таких данных точными эмоциональными метками — трудоёмкий процесс, который часто требует от экспертов интерпретации и классификации сложных эмоциональных состояний [12]. В то же время для эффективного распознавания эмоций необходимы крупномасштабные наборы данных, которые играют решающую роль в обучении нейронных сетей. Эти наборы данных позволяют моделям изучать сложные закономерности и вариации в выражениях эмоций в разных модальностях [16].

В первой части этого раздела мы рассмотрим классические подходы к классификации и описанию эмоций, для того чтобы иметь представление о том, каким образом аннотированы наборы данных. В двух следующих разделах рассмотрим наборы данных, аннотированные категориально и в многомерном пространстве эмоциональных состояний. Также будут отдельно рассмотрены наборы данных, содержащие тесты, используемые для разносторонней оценки алгоритмов и подходов классификации эмоциональных состояний.

1.1. Классификация эмоций

Как уже было сказано выше, описание и аннотация эмоциональных состояний являются сложными процессами [17]. Основные проблемы, которые при этом возникают, связаны с неопределенностью в числе классов эмоций и с неоднозначным описанием всех состояний заданным числом классов. Можно выделить два пути решения проблем описания и аннотирования эмоций: категориальный и многомерный подходы.

При категориальном подходе выбирается ограниченное число классов, которые могут в достаточной степени описывать все эмоциональные состояния человека. Например, в [18] предлагается различать две разные концепции базовых эмоций: как биологических примитивов и как психологических примитивов. В случае биологических примитивов в качестве основных признаков разделения эмоций принимаются биологические, такие как эволюционное происхождение, физиология и нейробиология возникновения этих эмоций. В случае психологических примитивов разделение базируется, в основном, на идее наиболее полного минимального описания эмоций, т.е. выбора такого ограниченного набора классов, который позволил бы статистически верно описывать все эмоциональные состояния человека. Примерами категориального подхода к классификации эмоций могут служить подход П. Экмана [19], который на основе подобия выражения и определения эмоций у людей из разных этнических групп выделял радость, удивление, печаль, гнев, отвращение, презрение, страх, и подход К. Изарда [20], выделяющий 11 базовых эмоций: радость, печаль, гнев, отвращение, удивление, страдание, волнение, презрение, смущение, стыд, вина на основе задействованных в выражении этих эмоций лицевых мышц. Такие подходы описания эмоций дают возможность их классифицировать, но зачастую не позволяют в достаточной степени разделять между собой близкие эмоции.

В случае многомерного подхода к описанию эмоций, предложенного В. Вундтом [21], возможно более точно описывать эмоции в силу их условной непрерывности. При этом предполагается,

Table 1. Comparison of existing multimodal categorical emotional datasets**Таблица 1.** Сравнение категориальных наборов мультимодальных эмоциональных данных

Название, год	Число участников	Число эмоциональных состояний	Количество модальностей	Модальности	Парадигма
ASCERTAIN [13], 2018	58	6 категорий	5	видео, ЭЭГ, КГР, акселерометр, движения лица, ЭКГ	видео
AMIGOS [23], 2021	40	6 категорий	4	3D видео, ЭЭГ, КГР, ЭКГ	видео, видео в группах
BP4D+ [16], 2016	140	7 категорий (экспертное аннотирование) 13 категорий (аннотирование испытуемыми)	7	3D видео, КГР, ЧДД, термография, ЭОГ, ЧСС, ФПГ	видео, диалог, задания
SEED-IV [24], 2018	15	4 категории	2	ЭЭГ, движения глаза	видео
ЭОГ — электроокулограмма; ЭЭГ — электроэнцефалограмма; ЭКГ — электрокардиограмма ЧСС — частота сердечных сокращений; ЧДД — частота дыхательных движений ФПГ — фотоплетизмограмма; КГР — кожно-гальваническая реакция					

что всю многогранность эмоций человека можно описать в трёхмерном пространстве, осями в котором являются измерения удовольствие — неудовольствие, расслабление — напряжение, возбуждение — успокоение, а каждая отдельная эмоция представляет собой точку в этом пространстве. Впоследствии исследователи предложили ещё несколько возможных комбинаций измерений в таких эмоциональных пространствах, следуя идее упрощения, но сохранения при этом достаточной описательной силы. В настоящее время преобладает вывод, что все эмоции достоверно и воспроизводимо отображаются в плоском эмоциональном пространстве активации и валентности [22]. Измерение валентности описывает положение состояния человека на шкале от максимального неудовольствия до максимального удовольствия, измерение активации на шкале от дремоты до бурного возбуждения.

Основываясь на приведённых способах классификации эмоций, можно так же разделить существующие наборы мультимодальных данных по способу аннотирования данных в них на категориальные и многомерные. Рассмотрим по отдельности такие наборы данных (табл. 1, 2).

1.2. Категориальные наборы данных

Одним из наиболее обширных категориальных корпусов эмоциональных данных, который включает в себя множество модальностей: периферические физиологические данные, видео с разных ракурсов, включая трёхмерное видео, записи с тепловизионных камер и аудио, является датасет, представленный в работе [16]. Набор содержит данные 140 испытуемых, каждый из которых принял участие в эксперименте, включающем в себя 10 эмоциональных заданий. Эмоциональные состояния испытуемых в течении 4 из этих заданий (целевые эмоции: счастье, смущение, страх и физическая боль) перекрёстно аннотированы 5 экспертами с использованием системы кодирования лицевых движений по 7 категориям. В итоге набор содержит 560 полностью аннотированных экспертами эпох данных, а так же 840 эпох данных, аннотированных на основании отрезков с наиболее ярким выражением эмоций длительностью 20 секунд. Также набор содержит описания испытуемыми переживаемых ими эмоций по 13 категориям для всех заданий. Существенным ограничением

при работе с данным корпусом эмоциональных состояний является объём данных — более 100 Гб данных для каждого испытуемого.

Набор данных ASCERTAIN [13] включает видео, аудио и физиологические данные. Этот датасет упрощает переход к реальному развертыванию систем, позволяя легко использовать модели в практических приложениях, т.к. периферические физиологические сигналы (ПФС) и электроэнцефалограмма (ЭЭГ) записаны с использованием доступных коммерческих носимых устройств. В отличие от других наборов данных, где используется сложное лабораторное оборудование, такие данные можно получать и в условиях реального применения разработанной системы классификации, используя те же устройства. В качестве стимула использовались 36 заранее отобранных видеороликов, реакция испытуемых на которые классифицировалась по 6 эмоциям. Суммарное число испытуемых составляло 58 человек.

Набор данных AMIGOS [23] интересен наличием двух парадигм: личной и групповой. В обоих случаях участники просматривали длинные и короткие видео-стимулы или находясь в одиночестве, или в группе. Набор содержит данные четырёх модальностей: 3D видео, электроэнцефалограмма, электрокардиограмма (ЭКГ), кожно-гальваническая реакция (КГР), аннотированные по шести эмоциональным классам. В эксперименте приняло участие 40 испытуемых.

Все наборы данных, рассмотренные в данном подразделе, доступны для некоммерческого использования по запросу, при заполнении пользовательского соглашения (EULA).

1.3. Многомерные наборы данных

Существует довольно много различных многомерных наборов эмоциональных данных. Все они в значительной мере различаются по количеству модальностей, числу испытуемых, пространствам эмоций, стимулам для возбуждения эмоций. Ниже представлены описания наиболее популярных и репрезентативных датасетов (табл. 2).

Набор данных CLAS [25], состоящий из данных четырёх модальностей — ЭЭГ, фотоплетизмограмма (ФПГ), КГР, акселерометр — представляет собой физиологические реакции пользователей как на неинтерактивные стимулы (видео, звук), так и на интерактивные стимулы (задания с малым временем, отведенным на ответ). Данные аннотированы в двух пространствах: пространстве эмоций возбуждения и валентности и пространстве концентрации и когнитивных способностей. В исследовании приняли участие 62 испытуемых, для каждого из которых был проведён эксперимент с одинаковым набором стимулов. Такой набор данных может быть крайне интересен для разработки человеко-машинных интерфейсов в силу сочетания достаточно простых для получения данных, не классовой, а непрерывной аннотации эмоций и ментального состояния и большого и разнообразного набора стимулов, вошедших в эксперимент.

Набор данных CASE [26] интересен своим способом аннотирования. Пользователи в процессе просмотра видео должны были обозначать своё эмоциональное состояние посредством джойстика в пространстве эмоций возбуждения и валентности. Набор данных содержит сигналы шести модальностей: ЭКГ, электромиограмма (ЭМГ), частота дыхательных движений (ЧДД), КГР, температура, ФПГ, записанные для 30 пользователей, которым демонстрировались 10 видеостимулов длительностью в среднем 2 минуты каждый.

Набор данных DEAP [27] включает в себя данные периферических физиологических сигналов и ЭЭГ для 32 испытуемых, а также видео мимики лица для 22 из них. Данные аннотированы в пространстве возбуждения и валентности, а также в метриках доминирования, симпатии и знакомства, по 9-балльной системе для всех метрик кроме знакомства, которое аннотировано по 5-балльной системе. Этот набор данных интересен наличием полноценного 32-канального ЭЭГ сигнала и большим количеством видео стимулов (40 шт.).

В ряду многомерных датасетов стоит упомянуть DECAF [28], содержащий мультимодальные данные 30 испытуемых, аннотированные по возбуждению, валентности и доминированию и по-

Table 2. Comparison of multimodal emotional datasets annotated in emotional spaces**Таблица 2.** Сравнение наборов мультимодальных эмоциональных данных аннотированных в эмоциональных пространствах

Название, год, доступность	Число участников	Метод аннотирования	Количество модальностей	Модальности	Парадигма
CLAS [25], 2019, (III)	62	ПВВ, ПКК	4	ЭЭГ, ФПГ, КГР, акселерометр	видео, задания
CASE [26], 2019, (I)	30	ПВВ	6	ЭМГ, КГР, ЭКГ, температура, ФПГ, ЧДД	видео
DEAP [27], 2012, (III)	32	ПВВ, 3 категории	8	ЭЭГ, КГР, ЧДД, ЭОГ, ЭМГ, ФПГ, температура, видео	видео
DECAF [28], 2007, (III)	30	ПВВ, 6 категорий	5	ик-видео, МЭГ, ЭМГ, ЭКГ, ЭОГ	видео
K-EmoCon [29], 2020, (II)	32	ПВВ, 18 категорий	6	ЭЭГ, видео, ЭКГ, акселерометр, ФПГ, температура	диалог
IEMOCAP [30], 2008, (III)	10	ПВВ, 8 категорий	3	текст, движения, видео	диалог
DREAMER [31], 2017, (II)	23	ПВВ	2	ЭЭГ, ЭКГ	видео
MOSEI [32], 2018, (I)	1000	(положительные, отрицательные)/ 5 уровней, 6 категорий	3	звук, видео, текст	монолог
Aff-Wild2 [33], 2019, (III)	554	ПВВ, 7 категорий	2	звук, видео	реальные условия
ПВВ — пространство эмоций возбуждения и валентности ПКК — пространство концентрации и когнитивных способностей ЭОГ — электроокулограмма; ЭЭГ — электроэнцефалограмма; ЭКГ — электрокардиограмма ЧСС — частота сердечных сокращений; ЧДД — частота дыхательных движений ФПГ — фотоплетизмограмма; КГР — кожно-гальваническая реакция (I) — доступен для некоммерческого использования (прямая загрузка) (II) — доступен при заполнении пользовательского соглашения на электронной платформе (III) — доступен по запросу при заполнении пользовательского соглашения					

лученные в экспериментах с теми же 40 видео-стимулами, как и в DEAP [27], а также 36 дополнительными видео-стимулами. Важным отличием DECAF от DEAP является то, что этот набор содержит помимо периферических физиологических сигналов данные магнитоэнцефалограммы (МЭГ) и видео лица в ближнем инфракрасном диапазоне, что позволяет сравнивать модальности ЭЭГ и МЭГ.

Набор данных K-EmoCon [29], включающий в себя мультимодальные данные 32 испытуемых, полученные в десяти 16-минутных диалогах, содержит периферийные физиологические сигналы, ЭЭГ и видео данные, аннотированные в пространстве возбуждения — валентности и по 18 кате-

гориям. Примечательной особенностью этого набора является выбор сенсоров, использованных для записи, а именно, данные были записаны с помощью носимых устройств.

Интересным для разработки бесконтактных методов распознавания эмоций является набор данных IEMOCAP [30], в котором содержатся такие модальности, как аудиовизуальные данные, мимика лица, а также текст. Несмотря на отсутствие физиологических сигналов типа ЭЭГ или ЭКГ, данный набор интересен своим количеством данных. В нём содержится более 12 часов записей 10 испытуемых актёров, которые разыгрывали эмоциональные состояния и диалоги. Данные разбиты на отрезки и аннотированы по восьми классам, а так же в пространстве валентность, активация, доминирование. Отличительной особенностью данного набора является использование меток в визуальных данных, как на лицах, так и на руках испытуемых, что позволяет более точно анализировать жестикуляцию и мимику.

Для тестирования и разработки методов распознавания эмоций на основе данных носимых датчиков физиологических сигналов, таких как ЭЭГ и ЭКГ, представляет интерес датасет DREAMER [31]. Набор содержит в себе данные 25 испытуемых, записанные в ходе экспериментов с 18 аудиовизуальными стимулами со средней продолжительностью 200 с. Данные аннотированы в пространстве валентность, активация, доминирование. Вследствие того, что запись выполнялась с помощью носимых сенсоров, данные характеризуются высоким уровнем шума. Тем не менее DREAMER, позволяет тестировать и разрабатывать методы для работы с данными носимых датчиков.

Набор данных Aff-Wild2 [33] содержит всего две модальности (звук и видео), но при этом интересен своим размером, и тем что содержит данные, полученные в максимально приближенных к реальным условиям. Датасет содержит 564 видео с 554 участниками. Видео взяты с видеохостинга (Youtube) и покaдрово аннотированы по 7 категориям (счастье, удивление, злость, отвращение, страх, грусть, нейтральное состояние) и в пространстве эмоций возбуждения и валентности непрерывными значениями от -1 до 1 . Аннотирование производилось усреднением мнений четырёх экспертов.

1.4. Наборы тестов

Отдельно стоит рассмотреть наборы данных содержащие в себе не только размеченные эмоциональные данные, но и основанную на них парадигму оценки (табл. 3). Такие наборы используются для сравнения и оценки алгоритмов классификации эмоций.

Стоит сказать про набор данных MERT [34], содержащий аудиовизуальные данные, и связанную с ним парадигму оценки, которые предназначены для оценки качества распознавания эмоций. Этот набор предоставляет стандартизированные тесты для оценки производительности систем распознавания эмоций и содержит данные 60 участников, по 30 оценок качества определения эмоций для 4 представленных в эксперименте парадигм.

Table 3. Comparison of emotional state classification test suites based on multimodal data

Таблица 3. Сравнение наборов тестов классификации эмоциональных состояний на основе мультимодальных данных

Название, год, доступность	Число участников	Число тестов	Количество модальностей	Модальности	Парадигма
MERT [34], 2009, (III)	60	4 теста	2	видео, звук	задания
UBFC-Phys [35], 2021, (II)	56	1 тест	4	видео, КГР, ФПГ, бесконтактное ФПГ	стресс-тест Трира
(II) — доступен при заполнении пользовательского соглашения на электронной платформе (III) — доступен по запросу при заполнении пользовательского соглашения					

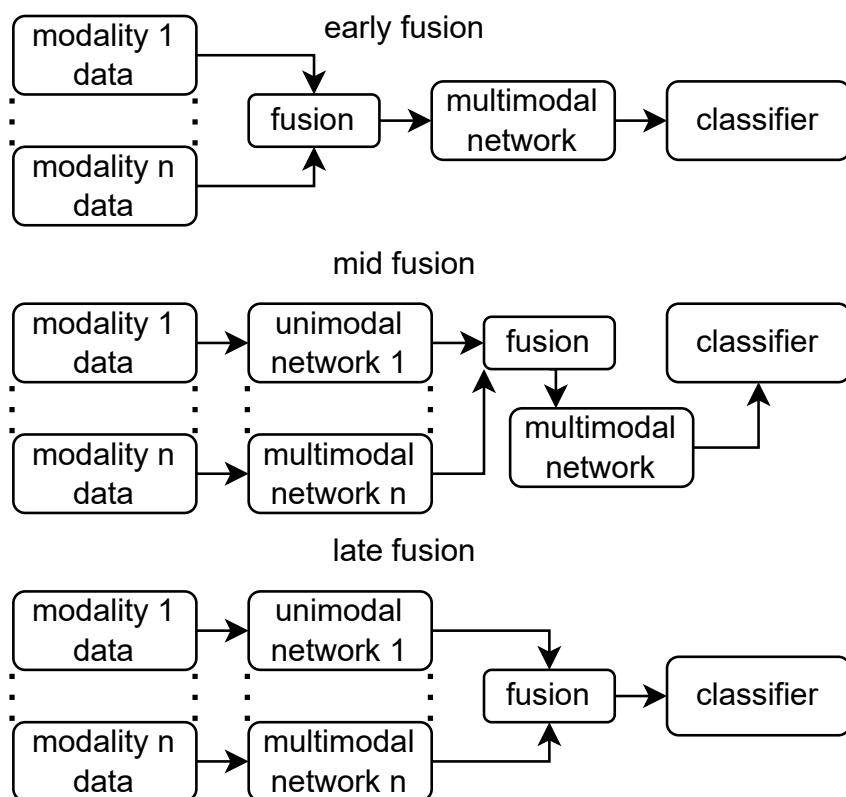


Fig. 1. Classification of multimodal methods by the stage at which occurs the fusion of modalities

Рис. 1. Классификация мультимодальных методов по этапу на котором происходит слияние модальностей

Набор данных UBFC-Phys [35] включает в себя видео и физиологические данные 56 испытуемых, полученные в ходе социального стресс-теста Трира. Интересной особенностью данного набора данных является использование в нем бесконтактной фотоплетизмограммы — бесконтактной замены метода измерения колебаний объёма крови в сосудах. Помимо этого записаны данные классической ФПГ. Такое сочетание позволяет использовать данный набор для улучшения определения уровня стресса у людей при помощи бесконтактных методов.

2. Методы машинного обучения для распознавания эмоций на основе мультимодальных данных

Методы распознавания эмоций могут быть дифференцированы между собой на основе различных признаков: по типу данных, которые они используют; по применяемым для классификации алгоритмам; по используемому подходу к классификации. Так же методы мультимодальной классификации можно подразделить по тому на каком этапе производится слияние модальностей данных на: методы с ранним слиянием, слиянием в середине, поздним слиянием и со смешанным типом слияния (рис. 1). В дальнейшем мы будем использовать данную классификацию для дополнительного подразделения мультимодальных методов. Перед тем, как перейти к рассмотрению

современных мультимодальных методов классификации, следует рассмотреть классические одно-модальные методы классификации.

При рассмотрении классических методов классификации эмоций, мы не будем подразделять их на методы для дискретной классификации и методы классифицирующие эмоции в пространстве эмоциональных состояний. В подавляющем большинстве случаев такое различие в типе классификации не влечёт за собой сколько-нибудь значительного изменения метода, лежащего в основе системы распознавания эмоций.

В первой части данного раздела мы рассмотрим одномодальные подходы к классификации эмоциональных состояний для самых распространённых модальностей данных. Далее в двух последующих разделах рассмотрим современные мультимодальные методы к классификации эмоций, а именно, методы дискретной классификации и методы классификации в пространстве эмоциональных состояний.

2.1. Классические подходы классификации эмоциональных состояний для отдельных модальностей

Одним из наиболее распространенных способов распознавания эмоций является анализ выражений лица. Этот метод основан на том, что определенные мышечные движения лица соответствуют различным эмоциональным состояниям. Существуют несколько подходов к распознаванию эмоций на основе анализа изображений лица: методы на основе выделения характеристик выражения лица и методы глубокого обучения. Метод выделения основных характеристик лица, таких как форма губ, глаз, бровей, заключается в выделении нескольких ключевых точек, положение которых используется для классификации эмоций. Одним из наиболее известных методов в таком подходе, является система кодирования лицевых действий (Facial Action Coding System, FACS), которая разбивает выражения лица на отдельные движения мышц (action units), из которых затем можно восстановить эмоцию [36]. Точность классификации (accuracy) эмоциональных состояний при помощи SVM на основе FACS может достигать 94.1 % [37]. Методы на основе глубокого обучения позволяют автоматизировать извлечение признаков из изображений лиц, а также производить классификацию по более полным данным, используя всё изображение для классификации, а не только признаки на основе ключевых точек. Обзор таких методов, достижения полученные при их применении и оставшиеся нерешёнными проблемы их применения рассматриваются в статье [38], рассмотренные в данной работе методы позволяют достигать качества классификации эмоций (accuracy) выше 95 %. Однако не стоит забывать, что данный подход требует наличия портретного видео лица, а наборы данных содержащие только выражения лиц значительно более просты в получении, следовательно, они часто содержат значительно большее число данных нежели мультимодальные наборы (например, набор данных Bosphorus [39] содержит 4666 2D и 3D изображений лиц 105 испытуемых).

Эмоции также могут быть распознаны по аудио сигналам, что особенно актуально в приложениях, где визуальная информация недоступна, например, в телефонных разговорах или голосовых помощниках. Основные подходы включают анализ спектральных признаков и глубокое обучение. Анализ спектральных признаков заключается в классификации эмоций на основе таких признаков как частота, громкость, темп и интонация. Одним из наиболее известных методов такого подхода является классификация на основе мел-кепстральных коэффициентов (Mel-Frequency Cepstral Coefficients, MFCCs) [40]. Подобно распознаванию лиц, в голосовых сигналах для распознавания эмоций активно применяются рекуррентные нейронные сети (Recurrent Neural Networks, RNN), включая их разновидности — долгую краткосрочную память (LSTM) и сети на основе механизмов внимания (Attention Mechanisms) [41] (accuracy по 7 категориям эмоций достигает 94 % за счёт использования RNN).

Table 4. Comparison of existing emotion classification methods based on unimodal data**Таблица 4.** Сравнение некоторых методов классификации эмоциональных состояний на основе данных отдельных модальностей

Набор данных	Модальность	Метод	Точность классификации (accuracy), %	Тип классификации
IEMOCAP	звук	SDT [46]	59.8	6 категорий
	видео	SDT [46]	41.5	6 категорий
	текст	SDT [46]	66.4	6 категорий
MELD	звук	SDT [46]	48.1	7 категорий
	видео	SDT [46]	48.1	7 категорий
	текст	SDT [46]	66.8	7 категорий
MOSEI	звук	SeMUL-PCD [47]	71.2	6 категорий
	видео	SeMUL-PCD [47]	72.1	6 категорий
	текст	SeMUL-PCD [47]	84.5	6 категорий
SEED-IV	ЭЭГ	BiHDM [48]	74.4	4 категорий
	взгляд	MFFNN [49]	64.4	4 категорий
DEAP	ЭЭГ	NeuroSense [50]	73.5	ПБВ*
ПБВ* – ПБВ, упрощённое до двух задач бинарной классификации				

Текстовая информация также может иметь значительную эмоциональную окраску. Подходы анализа текста для распознавания эмоций также можно подразделить на классические и использующие глубокое обучение. К классическому подходу можно отнести методы на основе мешка слов и n-грамм — методы на основе частотного анализа слов и фраз, где текст представлялся в виде набора слов с соответствующими частотами встречаемости этих слов. Для классификации эмоций на основе признаков полученных из анализа текста применялись такие алгоритмы, как наивный Байесовский классификатор или SVM [42] (accuracy бинарной классификации эмоций 86.2 %). Современные методы глубокого обучения используют нейронные сети трансформеры такие, как BERT и GPT, которые способны учитывать контекст и сложные взаимосвязи слов в предложении. Эти модели обучаются на больших корпусах текстов и могут предсказывать эмоциональные категории текста с высокой точностью [43] (accuracy классификации эмоций по 7 классам для набора данных MELD [44] 61.9 %).

Так же классификация эмоций может быть произведена на основе физиологических данных, таких как: электроэнцефалограмма, электрическая активность кожи, вариабельность сердечного ритма и дыхания, а также их комбинаций. Подходы классификации таких данных так же развивались от классификации отдельных признаков сигнала к использованию методов глубокого обучения объединяющих данные нескольких модальностей [45] (accuracy классификации 7 эмоциональных состояний на основе КГР: 57.1 %, на основе КГР, жестов, выражений лица: 78.3 %).

В таблице 4 рассмотрены методы классификации на основе отдельных модальностей данных, для наборов данных использованных в рассматриваемых далее мультимодальных подходах.

2.2. Современные методы дискретной мультимодальной классификации эмоций

Современные системы, использующие мультимодальные подходы, комбинируя различные типы данных для более точного распознавания эмоций, обычно применяют ансамблевые модели или архитектуры с несколькими входными кодировщиками, что позволяет учитывать разнообразные признаки и улучшать точность классификации [51]. Модели, объединяющие несколько модальностей, во многих случаях позволяют получить лучшие результаты распознавания эмоций по сравнению с одномодальными аналогами [52, 53].

Применяемые при этом методы глубокого обучения включают различные нейросетевые архитектуры: нейронные сети доверия, глубокие свёрточные нейронные сети (DCNN), сети долгой краткосрочной памяти (LSTM) [54], и их комбинации [55].

Рассматриваемые в данном разделе методы разбивают эмоции на заранее заданные классы. Такой подход позволяет упростить как аннотирование данных, так и саму задачу классификации. Однако такое упрощение делает невозможным применение данных методов для распознавания эмоций, не относящихся к исходному набору классов. Архитектуру мультимодальных методов классификации эмоций можно условно разделить на 3 основные части: блок, выделяющий признаки из данных исходных модальностей, блок, отвечающий за слияние признаков из разных модальностей и выделение признаков из мультимодальных данных, и блок итоговой классификации. Наиболее интересны различия представленных методов в части исходной обработки данных, слияния модальностей и наборе используемых модальностей.

2.2.1. Современные методы дискретной мультимодальной классификации эмоций с ранним слиянием модальностей

Интересным подходом к классификации эмоций на основе мультимодальных данных, является COGMEN (Graph Neural Network based Multimodal Emotion recognitionN) [56] — метод, совмещающий в себе графовую нейронную сеть и архитектуру трансформер (рис. 2). В данной модели слияние данных различных модальностей происходит на раннем этапе вместе с выделением признаков данных при помощи кодировщика на основе архитектуры трансформер. На следующем этапе данные проходят через модуль формирования графа, который позволяет учитывать связи между дикторами в диалоге, а также связи между высказываниями каждого говорящего. Далее, при помощи реляционной графовой свёрточной нейронной (RGCN) сети выделяются сложные признаки данных, которые попадают на вход модуля внимания, отвечающего за итоговое слияние модальностей. На основе обработанных таким образом данных с помощью полносвязных слоев происходит итоговая классификация эмоциональных состояний. Представленный метод позволяет использовать сразу три модальности: аудио, видео и текст, а так же контекст диалога между пользователями, демонстрируя при этом отличные результаты классификации (F1-мера для 4 эмоциональных классов набора данных IEMOCAP равна 84.5 %). Наиболее интересным решением в данном методе является применение графовых нейронных сетей для обработки контекста в диалоге.

2.2.2. Современные методы дискретной мультимодальной классификации эмоций со слиянием модальностей в середине

В работе [57] разработана мультимодальная модель глубокого обучения, которая использует изображения лиц и текстовые данные, описывающие контекст, для того чтобы классифицировать эмоции персонажей корейского сериала «Мисэн: Неполное» по семи категориям: ярость, отвращение, ужас, радость, нейтральность, грусть и удивление. Представленная архитектура достаточно проста и состоит из свёрточных слоёв для выделения признаков из изображений, кодировщика текстовых данных (Glove/ELMo) и полносвязных слоёв для итоговой классификации. Объединение модальностей происходит при помощи простой конкатенации выходов ветвей модели. В работе получены результаты, демонстрирующие улучшение качества классификации эмоциональных состояний мультимодальной моделью (видео и текст, 7 категорий F1: 50 %) в сравнении с одномодальной (видео, F1: 32 %), даже при использовании достаточно простой архитектуры и легкодоступных модальностей данных.

В работе [9] предложен метод объединения модальностей для архитектуры на основе трансформера. Разработанный метод позволяет объединить 3 модальности (изображение, текст и звук) для классификации эмоций. Выделение признаков из данных происходит при помощи свёрточных нейронных сетей для звука и изображений, а для выделения признаков из текста применяется пред-

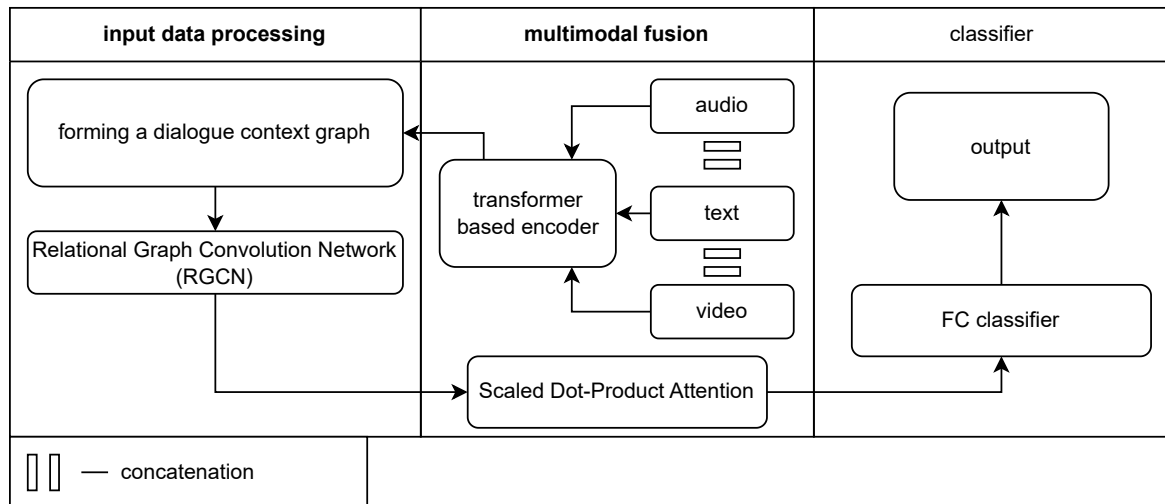


Fig. 2. Schematic of discrete multimodal emotion classification architecture with early fusion

Рис. 2. Схема архитектуры дискретной мультимодальной классификации эмоций с ранним слиянием модальностей

варительно обученная нейронная сеть с архитектурой подобной BERT (ALBERT). Слияние модальностей происходит с помощью модуля на основе архитектуры трансформера (рис. 3). Для итоговой классификации данных применяется архитектура на основе трансформера и полносвязного слоя, что позволяет применять обучаемые отдельно для каждого видео из набора данных эмбединги эмоций. Исследователям удалось продемонстрировать, что использование архитектуры трансформера для слияния модальностей позволяет улучшить качество классификации (F1: 60.9 %), в сравнении с применением конкатенации для слияния (F1: 57.2 %). Ключевыми особенностями метода, представленного в работе [9], являются:

- метод слияния модальностей на основе трансформера, позволяющий выделять более сложные мультимодальные признаки из данных;
- устройство классифицирующей части алгоритма, позволяющее использовать все данные, полученные из отвечающего за слияние модальностей модуля, а не только классифицирующий токен (CLS), а также дающее возможность применять эмбединги эмоций для каждого видео.

Развитие идеи применения графовых нейронных сетей для учёта контекста в классификации эмоций в диалоге продемонстрировано в работе [58]. Для выделения признаков данных из отдельных модальностей в данном подходе использовались архитектуры RoBERTa, openSMILE и 3D-CNN для текстовой, звуковой и видео модальностей соответственно. Далее полученные признаки данных поступали на вход модуля формирования графа, отвечающего за слияние модальностей и учёт контекста. Полученные графовые представления мультимодальных данных попадали на вход архитектуры, состоящей из двух ветвей графовой нейронной сети Фурье (Fourier Graph Neural Network). Эти ветви обладают различными частотными характеристиками: одна обрабатывает признаки низкой частоты, а вторая — признаки более высокой частоты. Такой подход позволил не только учитывать медленное и быстрое изменение эмоциональных состояний, но и применить метод самообучения — контрастивное обучение (Contrastive Learning). Полученные таким образом признаки данных использовались для итоговой классификации. Авторам удалось продемонстрировать высокое качество классификации на наборе данных IEMOCAP (ассураку для 6 эмоциональных классов равна 73.8 %, что является очень хорошим результатом для указанного датасета). Наиболее интересным новшеством представленного в данной работе метода, является применение метода контрастивно-

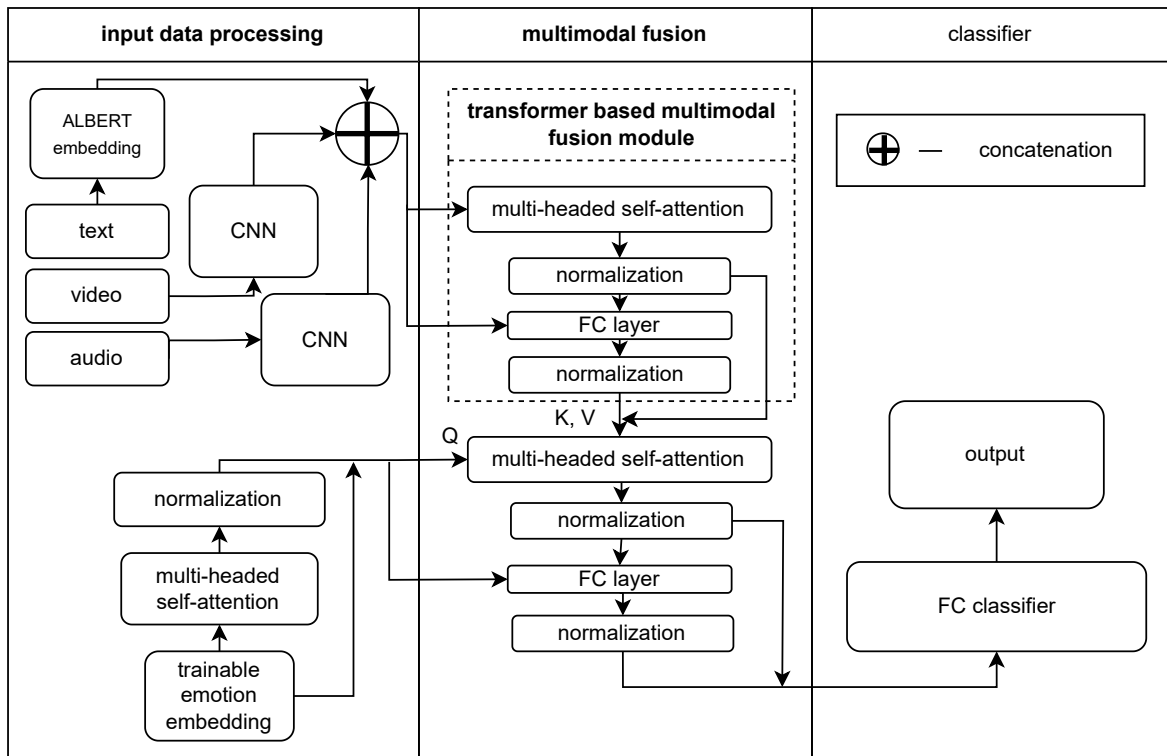


Fig. 3. Schematic of discrete multimodal emotion classification architecture with middle fusion

Рис. 3. Схема архитектуры дискретной мультимодальной классификации эмоций со слиянием модальностей в середине

го обучения для графовых нейронных сетей Фурье. В сравнении с представленным ранее методом на основе графовых нейронных сетей (COGMEN [56]), авторам удалось продемонстрировать улучшение качества классификации (F1: 73.9 %, для COGMEN F1: 67.6 % (набор данных: IEMOCAP, 6 классов эмоций)), по средствам оптимизации метода обучения и добавления кодировщиков, выделяющих признаки из отдельных модальностей.

2.2.3. Современные методы дискретной мультимодальной классификации эмоций с поздним слиянием модальностей

Возможность использования предварительно обученных «BERT-подобных» архитектур для выделения сложных признаков из звуковых и текстовых данных в виде токенов показана в работе [59]. Полученные таким образом токены используются для мультимодального распознавания эмоций в речи. Исследователи рассмотрели несколько подходов слияния модальностей: с использованием механизма внимания (Co-Attentional Fusion) и простое слияние посредством конкатенации токенов (Shallow-Fusion). Они продемонстрировали, что базовый механизм слияния Shallow-Fusion превосходит механизм на основе внимания, упрощает общую структуру нейронной сети и улучшает качество классификации. Схема предложенной в этой статье архитектуры приведена на рис. 4а.

В [11] описан подход (рис. 4б) для получения признаков звуковых и текстовых данных для классификации эмоциональных состояний. С помощью технологии полосовой фильтрации (слой SincNet) и нейронной сети извлекались акустические характеристики из необработанного звука. Для текстовой модальности данных генерировался набор представлений данных при помощи глобальных векторов (GloVe embeddings), после чего это представление обрабатывалось архитектурой, содержа-

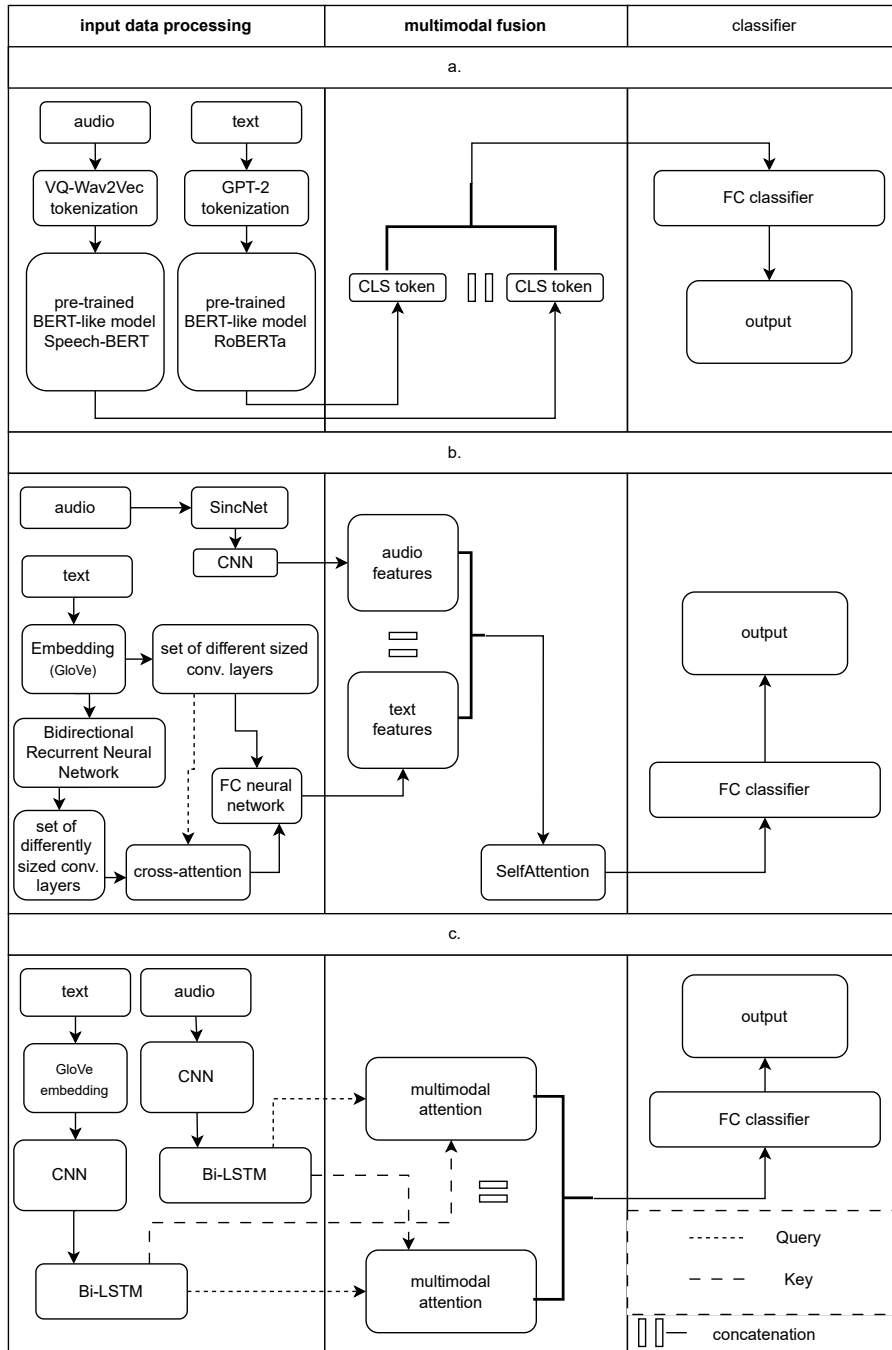


Fig. 4. Schematic of discrete multimodal emotion classification architecture with late fusion (a — from [59], b — from [11], c — from [60])

Рис. 4. Схемы архитектур дискретной мультимодальной классификации эмоций с поздним слиянием модальностей (a — архитектура из [59], b — архитектура из [11], c — архитектура из [60])

щей две параллельные ветви. Одна из этих ветвей представляла собой рекуррентную сеть, а вторая — свёрточную сеть с набором фильтров разного размера для захвата N-грамм различной длины. Также выход свёрточной сети использовался в модуле перекрестного внимания с ветвью, представленным рекуррентной сетью. Выход текстовой части сети обрабатывался сетью классификации эмоций для текстовой модальности (T-DNN). Результат работы сетей классификации текстовой и звуковой модальностей объединялся при помощи механизма внимания Self-Attention. Наиболее интересными решениями представленного метода являются:

- применение слоя SincNet в роли набора обучаемых фильтров, который позволяет использовать аудио данные (голос) без предварительной обработки;
- применение двунаправленной рекуррентной сети параллельно со свёрточной нейронной сетью для выделения признаков из текстовых данных. Такое решение позволяет наиболее эффективно выделять контекстуальные признаки и признаки на основе N-грамм.

Модель классификации эмоциональных состояний на основе текста и звука представленная в работе [60], состоит из свёрточных слоёв для извлечения простых признаков данных и двунаправленных сетей долгой краткосрочной памяти (Bi-LSTM) для извлечения более сложных признаков (рис. 4с). Примечательным в данном подходе является способ слияния текстовой и звуковой модальностей на основе двух отдельных модулей мультимодального внимания. Разработанный в данной работе подход демонстрирует высокое качество классификации эмоциональных состояний (ассигасу: 72.8 %) при достаточно простой архитектуре, отвечающей за извлечение признаков данных. В работе также приведено качество классификации на основе данных только одной модальности, а именно, точность (ассигасу) 60.7 % на звуковых данных и 65.9 % на текстовых. Приведенные результаты могут свидетельствовать о высокой эффективности представленного в данной работе механизма слияния модальностей. Стоит заметить, что продемонстрированное в данной работе качество классификации (ассигасу: 72.8 %), ниже нежели качество классификации продемонстрированное в работе [59] (ассигасу: 75.4 %), даже при использовании вероятно более эффективного метода слияния модальностей на основе внимания. Это может свидетельствовать о большей эффективности BERT-подобных архитектур в качестве кодировщиков данных исходных модальностей в сравнении со свёрточными нейронными сетями (CNN).

Существует довольно много способов объединения модальностей с учетом особенностей каждой из них, в том числе, метод MFFNN (Multimodal Feature Fusion Neural Network) [61], использующий для этой цели модуль на основе внимания. Данный модуль позволяет вычленять наиболее значимые для классификации признаки каждой модальности, а так же выравнивать временные характеристики модальностей. Разработанный в работе метод позволяет классифицировать эмоциональные состояния на основе данных ЭЭГ и взгляда. Архитектура представленной нейронной сети состоит из свёрточных слоёв для выделения признаков из данных, слоя слияния модальностей на основе мультимасштабное внимания (MS Fusion) и классификатора на основе полносвязного слоя. Наибольшего внимания заслуживает именно применённый метод внимания, позволивший эффективно объединить сложные с точки зрения временных характеристик данные ЭЭГ и взгляда. Авторам удалось продемонстрировать улучшение точности классификации эмоций при использовании мультимодальных данных в сравнении с одномодальными, а также получить итоговое качество классификации, превосходящее качество классификации иных современных методов.

2.2.4. Метод дискретной мультимодальной классификации эмоций со смешанным слиянием модальностей

В работе [62] представлена мультимодальная модель машинного обучения с механизмом внимания (MMAN) (рис. 5), для распознавания эмоций на основе 3 модальностей: аудио, текст и изображение. Одной из особенностей предложенного метода является механизм слияния модальностей, объединяющий 2 подхода. Входные данные обрабатываются отдельными для каждой модальности

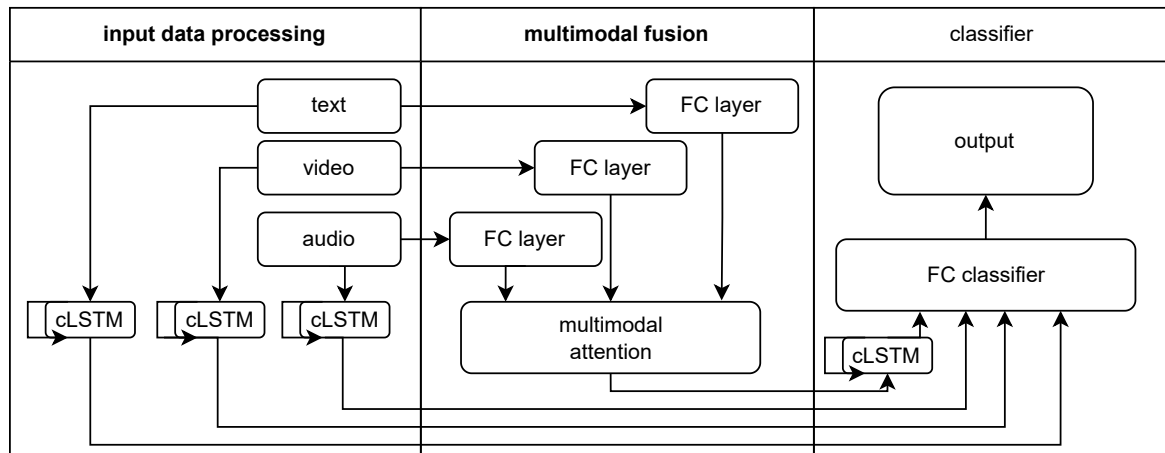


Fig. 5. Schematic of discrete multimodal emotion classification architecture with mixed fusion

Рис. 5. Схема архитектуры дискретной мультимодальной классификации эмоций со смешанным слиянием модальностей

сти сетями на основе архитектуры контекстной долгой краткосрочной памяти (cLSTM), а также сетью cLSTM-MMA объединяющей данные трёх модальностей. После чего результаты всех 4 сетей подаются на вход полносвязного слоя, отвечающего за итоговое слияние модальностей и классификацию эмоций. Архитектура cLSTM-MMA включает в себя модуль внимания между модальностями и cLSTM. Тесты показывают, что при классификации речевых эмоций визуальные и текстовые данные могут значительно повысить качество классификации. Наиболее интересным в данном подходе является применение отдельных сетей долгой краткосрочной памяти для получения признаков из каждой модальности отдельно, а также совместно с модулем внимания для мультимодальных данных. Такой подход позволяет использовать для итоговой классификации как признаки, полученные из отдельных модальностей, так и признаки из их объединения.

2.2.5. Современные методы оптимизации процесса обучения нейронных сетей дискретной мультимодальной классификации эмоций

Для улучшения качества классификации эмоциональных состояний на основе мультимодальных данных важна не только разработка методов выделения признаков из отдельных модальностей и методов слияния модальностей, но и методы обучения мультимодальных моделей. В исследованиях [63, 64] продемонстрировано, что добавление новой модальности может приводить к ухудшению качества классификации модели в сравнении с методами, основанными на отдельных модальностях. Авторы в данных работах демонстрируют, что причиной такого ухудшения является дисбаланс обучения мультимодальной архитектуры в случае, когда конкуренция между модальностями на этапе обучения ухудшает качество обучения для отдельных модальностей. Одним из вариантов решения данной проблемы является применение методов балансировки обучения [65, 66], однако такие методы требуют вычисления функции ошибки для каждой модальности отдельно. Данная задача может быть решена несколькими способами:

- маскированием частей нейронной сети, отвечающих за отдельные модальности [65], но такой подход влечёт за собой увеличение вычислительных затрат и накладывает ограничения на структуру мультимодальной модели;
- добавлением отдельных классификаторов для каждой модальности на этапе обучения.

В работе [66] предложен метод решения указанной проблемы путем дополнения мультимодальной архитектуры отдельными классификаторами для каждой из модальностей (звук и видео)

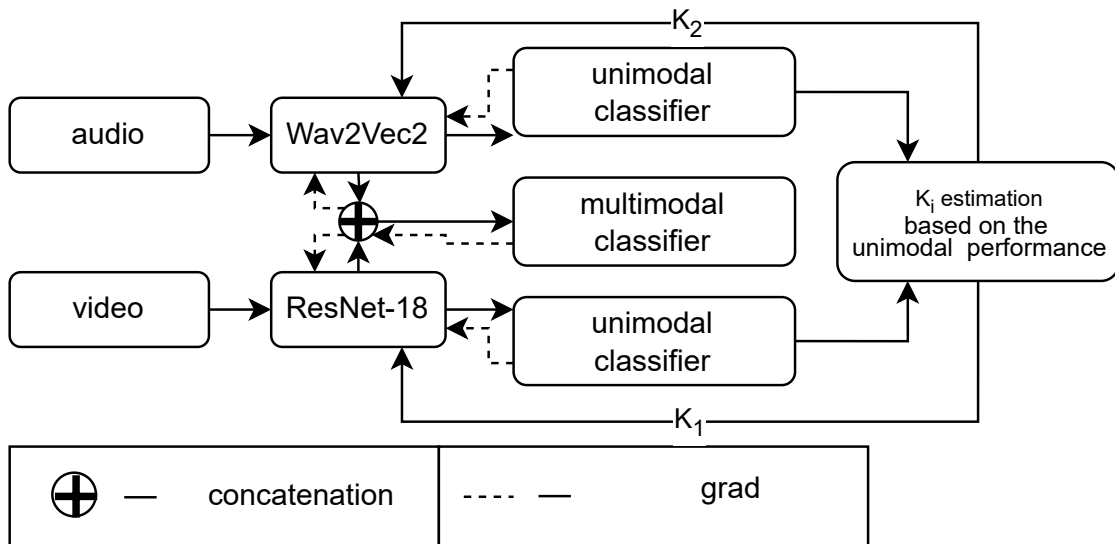


Fig. 6. Schematic diagram of the training balancing architecture of the multimodal emotion classification neural network

Рис. 6. Схема архитектуры балансировки обучения мультимодальной нейронной сети классификации эмоций

(см. рис. 6), используемыми в процессе обучения модели для вычисления функции потерь и качества классификации отдельных модальностей. В качестве функции потерь использовалась сумма функций потерь (перекрёстной энтропии) для каждой отдельной модальности и мультимодального классификатора. Качество классификации классификаторов каждой модальности использовалось для выбора K_i — коэффициентов балансировки скорости обучения модулей отвечающих за выделение признаков из отдельных модальностей. Такой подход позволил исследователям сбалансировать мультимодальное обучение, избегая недо- и переобучения отдельных частей нейронной сети.

Авторы использовали архитектуру, состоящую из ResNet-18 и Wav2Vec2 [67], для выделения признаков из видео и звуковых данных и два отдельных метода слияния модальностей — на основе полносвязного слоя и на основе архитектуры трансформер. Им удалось продемонстрировать улучшение качества классификации эмоциональных состояний и снижение дисбаланса качества обучения частей архитектуры, отвечающих за отдельные модальности, при использовании предложенного метода обучения (accuracy: 80 %) в сравнении с классическим подходом к обучению (accuracy: 70 %).

Обобщение результатов для рассмотренных алгоритмов дискретной классификации эмоций приведено в таблице 5.

2.3. Методы мультимодальной классификации эмоций в пространстве эмоциональных состояний

Методы из данного раздела классифицируют эмоции в пространстве эмоций возбуждения и валентности, такой подход усложняет аннотирование данных при подготовке набора данных и саму задачу классификации, превращая её в задачу регрессии. При этом классификация в пространстве возбуждения и валентности может позволять классифицировать эмоции, не содержащиеся в обучающем наборе данных. Часто для упрощения задачи классификации исследователи заменяют непрерывные значения в пространстве эмоций возбуждения и валентности (ПВВ) на несколько дискретных значений для каждой из шкал [68, 69]. Такое упрощение позволяет свести задачу регрессии к задаче мультиклассовой или даже бинарной классификации для шкал возбуждения и валентно-

Table 5. Comparison of existing methods for discrete emotion classification based on multimodal data**Таблица 5.** Сравнение существующих методов дискретной классификации эмоций на основе мультимодальных данных

Базовая архитектура/ метод слияния модальностей/ этап слияния/ классификатор	Набор данных	Модальности	Кол-во классов	Качество классификации	Ссылка, год
RGCN/ transformer/ ранний/ FC classifier	IEMOCAP, MOSEI	звук, видео, текст	4/6	IEMOCAP: F1: 84.5 % / 67.6 % MOSEI: F1: 76.8 % / 79.3 %	[56], 2022
CNN/ конкатенация/ средний/ FC classifier	свой	звук и видео	7	F1: 50 %	[57], 2020
transformer, CNN, BERT/ transformer/ средний/ FC classifier	IEMOCAP, MOSEI	видео, звук, текст	6	accuracy: IEMOCAP: 85.9 % MOSEI: 67.8 %	[9], 2023
FourierGNN, FC/ GNN/средний/ FC classifier	IEMOCAP, MELD	звук, видео, текст	6	accuracy: IEMOCAP: 73.8 % MELD: 68.1 %	[58], 2025
BERT-like/ конкатенация/ поздний/ FC classifier	IEMOCAP	звук и текст	4	accuracy: 75.4 %	[59], 2020
BiRNN, CNN/ attention/ поздний/ FC classifier	IEMOCAP	звук и текст	4	accuracy: 79.2 %	[11], 2020
Bi-LSTM, CNN/ attention/ поздний/ FC classifier	IEMOCAP	звук и текст	4	accuracy: 72.8 %	[60], 2020
CNN/attention/ поздний/ FC classifier	SEED-IV	ЭЭГ и взгляд	4	accuracy: 87.3 %	[61], 2023
cLSTM/ attention/ смешанный/ FC classifier	IEMOCAP	звук, видео и текст	4	accuracy: 73.9 %	[62], 2022

сти. Такое упрощение снижает итоговое число классов эмоций, которые могут быть однозначно классифицированы, однако позволяет сохранить возможность разделять между собой классы эмоций, не содержащиеся в обучающем множестве. Архитектуры, лежащие в основе рассмотренных в данном разделе методов, можно разделить на те же условные части, что и в предыдущем разделе: выделение признаков данных, слияние модальностей, классификация. Наиболее существенными

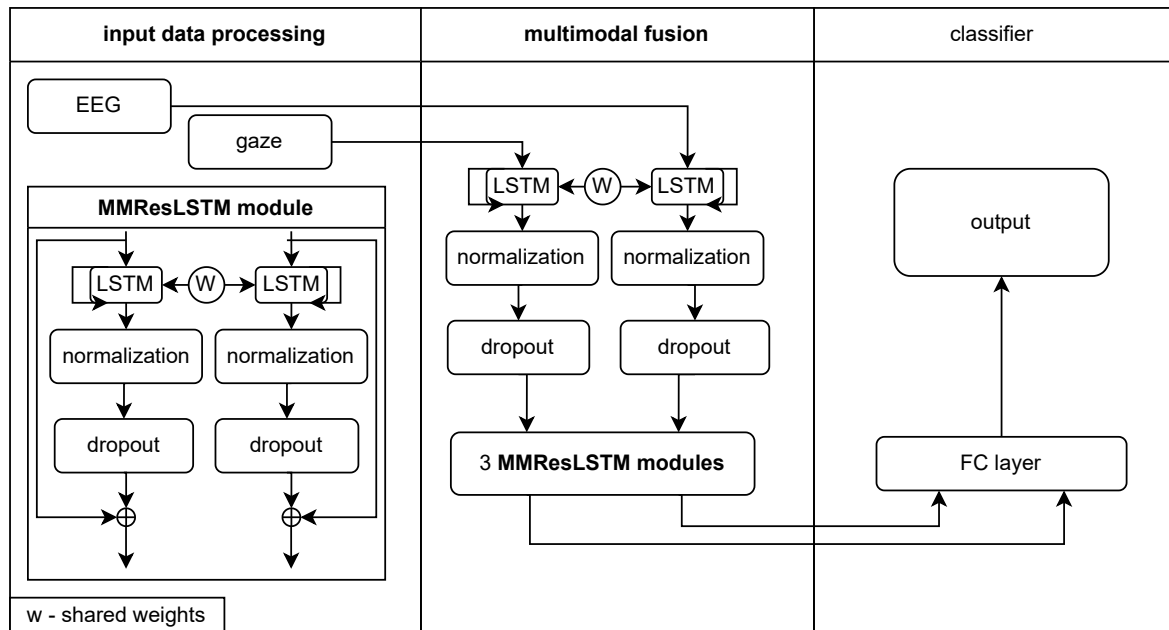


Fig. 7. Schematic of multimodal emotion classification architecture in arousal–valence space with early fusion

Рис. 7. Схема архитектуры мультимодальной классификации эмоций в пространстве возбуждения и валентности с ранним слиянием модальностей

различиями в рассматриваемых методах являются используемые модальности, методы выделения признаков из данных и методы слияния модальностей.

2.3.1. Современные методы мультимодальной классификации эмоций в пространстве возбуждения и валентности с ранним слиянием модальностей

Мультимодальная нейронная сеть (MMResLSTM) (рис. 7) для улучшения распознавания эмоций в ПВВ, упрощённом до двух задач бинарной классификации, с использованием данных ЭЭГ и движения глаза представлена в работе [69]. Используя общие веса между модальностями, MMResLSTM фиксирует корреляции между модальностями, как пространственные, так и временные, для эффективного представления признаков высокого уровня. Качество классификации (ассурасу) предложенного алгоритма в пространстве возбуждения и валентности на данных набора DEAP достигает 92 %. Такой подход позволяет улучшить качество классификации в сравнении с методом [70] (ассурасу 83.5 %) на основе подобной архитектуры (Bi-LSTM), но использующим другой метод слияния модальностей (конкатенация). Следовательно, использование архитектуры MMResLSTM как для выделения признаков из данных, так и для слияния модальностей при помощи общих весов является интересным решением проблемы слияния модальностей.

2.3.2. Современные методы мультимодальной классификации эмоций в пространстве возбуждения и валентности со слиянием модальностей в середине

Применение мультимодального подхода для непрерывного представления эмоций в пространстве возбуждения и валентности на основе визуальных и звуковых данных продемонстрировано в работе [71]. Для извлечения признаков из данных применялись свёрточные нейронные сети, а для слияния двух модальностей и использования контекста — двухслойная архитектура LSTM (рис. 8). Авторам удалось добиться относительно высокого качества предсказания эмоциональных

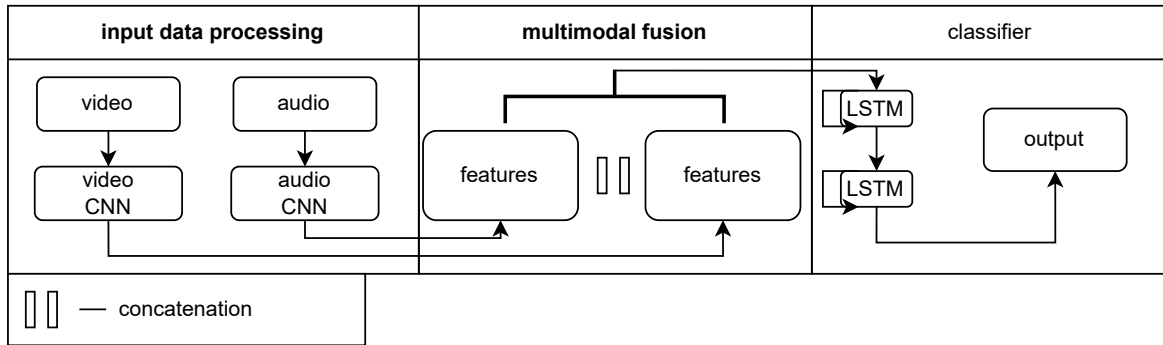


Fig. 8. Schematic of multimodal emotion classification architecture in arousal — valence space with middle fusion

Рис. 8. Схема архитектуры мультимодальной классификации эмоций в пространстве возбуждения и валентности со слиянием модальностей в середине

состояний в ПВВ (коэффициент корреляции конкордации — $\rho_c = 0.789$) при использовании достаточно простой архитектуры нейронной сети. Также в работе приведены данные, наглядно демонстрирующие улучшение качества классификации с использованием мультимодального подхода по сравнению с подходами на основе отдельных модальностей (видео $\rho_c = 0.426$, звук $\rho_c = 0.556$).

Метод, занявший первое место в шестом конкурсе Affective Behavior Analysis in-the-Wild [33], представлен в работе [72]. Исследователи применили предварительно обученную нейронную сеть Wav2Vec2 [67] для выделения признаков из аудио данных и нейронную сеть на основе маскированного автокодировщика (MAE) для выделения признаков видео данных. Для слияния признаков данных двух модальностей использовался модуль на основе архитектуры трансформера. Отличительной особенностью рассмотренного подхода является метод, использованный для обработки временных признаков данных. В нем элементами последовательности данных, поступающими на вход метода слияния модальностей, являлись признаки, полученные не для каждого отдельного кадра, а выделенные для последовательности кадров видео заданной длины. Для итогового покадрового предсказания уровней возбуждения и валентности применялось сглаживание. Коэффициент корреляции конкордации на тестовом наборе данных (Aff-Wild2), продемонстрированный в данной работе, равен $\rho_c = 0.677$, что выше $\rho_c = 0.619$ представленного в работе [73]. Основными отличиями данных методов являются метод слияния модальностей и подход к представлению временной последовательности данных, что может свидетельствовать о эффективности применения метода слияния на основе архитектуры трансформера, и использования сглаживания для предсказания представления эмоции в ПВВ.

2.3.3. Современные методы мультимодальной классификации эмоций в пространстве возбуждения и валентности с поздним слиянием модальностей

В статье [68] предложен метод мультимодальной классификации эмоций в ПВВ упрощённом до двойной задачи бинарной классификации на основе бимодального глубокого автокодировщика BDAE (Bimodal Deep AutoEncoder), и продемонстрировано улучшение качества классификации по сравнению с другими существовавшими на тот момент методами. Разработанный метод на данных ЭЭГ и движений глаза превосходит по качеству классификации методы на отдельных модальностях, а также ансамбль таких методов. Качество классификации предложенного алгоритма в пространстве возбуждения и валентности на данных набора DEAP составляет 83 %. Архитектура данного метода представлена на рис. 9а.

В работе [70] авторы используют преимущества модели LSTM в составлении сложных корреляций и учете контекста для отдельных модальностей, после чего классифицируют эмоции на основе

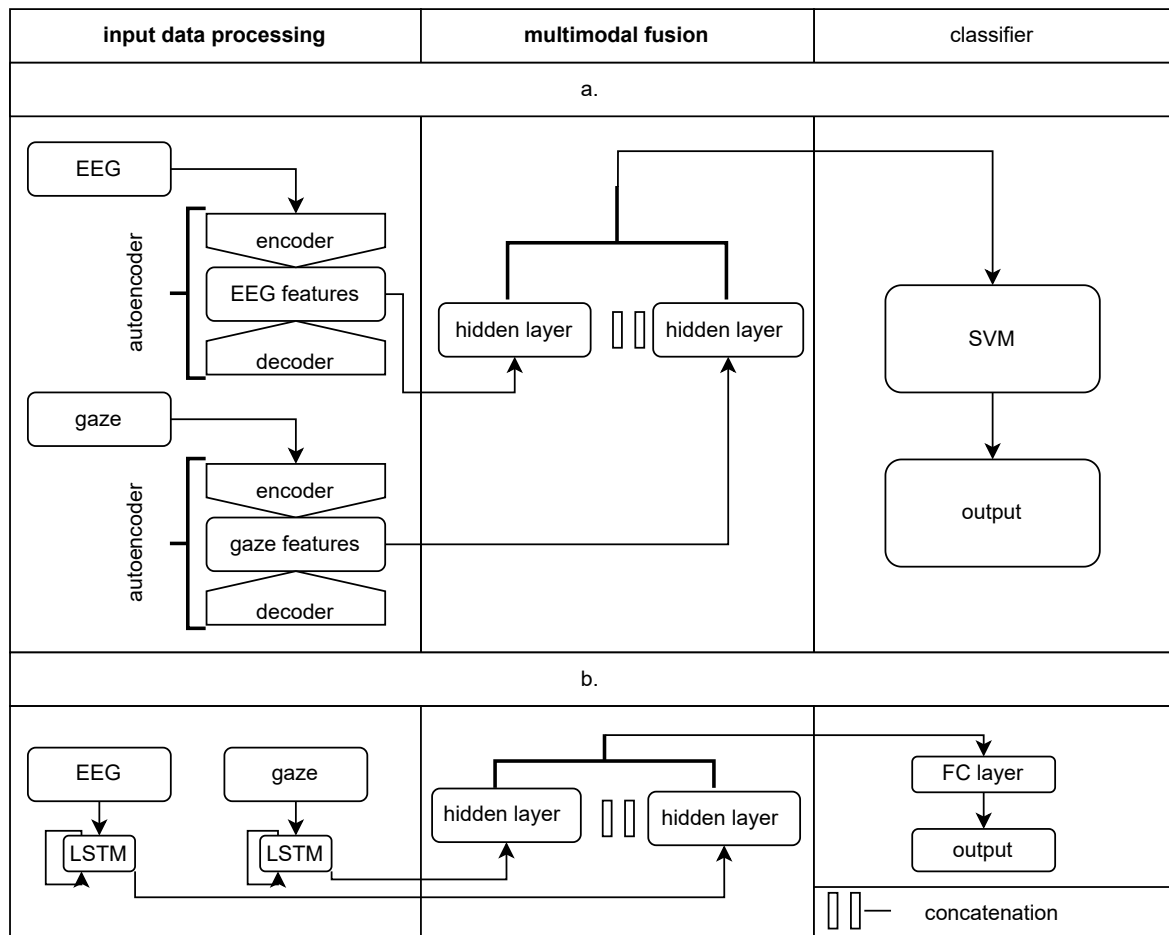


Fig. 9. Schematics of multimodal emotion classification architectures in arousal—valence space with late fusion

Рис. 9. Схемы архитектур мультимодальной классификации эмоций в пространстве возбуждения и валентности с поздним слиянием модальностей

полученных с помощью LSTM сложных признаков данных (рис. 9b). Качество классификации в пространстве возбуждения и валентности, упрощённом до двойной задачи бинарной классификации, на данных из набора DEAP достигло 83.5 %. Наиболее интересным результатом данной работы стало то, что исследователям удалось продемонстрировать качество классификации, превосходящее более сложные модели, используя достаточно простую архитектуру как для выделения признаков, так и для слияния модальностей данных.

В статье [73] исследователи продемонстрировали метод мультимодальной классификации эмоциональных состояний в пространстве возбуждения и валентности (непрерывными значениями от -1 до 1). Используемая в данной работе архитектура состоит из:

- методов извлечения признаков данных: кодировщика на основе архитектуры VGG-Net для данных аудио, кодировщика на основе Resnet для видео и кодировщика на основе архитектуры BERT для текстовых данных;
- метода слияния модальностей на основе модулей совместного внимания (Joint Attention);
- классификатора на основе многослойного перцептрона.

Table 6. Comparison of existing methods for emotion classification in emotional state space based on multimodal data**Таблица 6.** Сравнение существующих методов классификации эмоций в пространстве эмоциональных состояний на основе мультимодальных данных

Базовая архитектура/ метод слияния модальностей/ этап слияния/ классификатор	Набор данных	Модальности	Тип классификации	Качество классификации	Ссылка, год
ResLSTM/shared weights/панний/ FC classifier	DEAP	ЭЭГ и взгляд	ПБВ*	accuracy: 92 %	[69], 2019
CNN + LSTM/ конкатенация/ средний/LSTM	RECOLA	видео и звук	ПБВ	$\rho_c = 0.789$	[71], 2017
CNN, MAE/ transformer/ средний/ FC classifier	Aff-Wild2	видео и звук	ПБВ	$\rho_c = 0.677$	[72], 2023
BDAE/ конкатенация/ поздний/SVM	SEED-IV, DEAP	ЭЭГ и взгляд	SEED-IV: 4 кат. DEAP: ПБВ*	accuracy: SEED-IV: 91 % DEAP: 83.3 %	[68], 2016
Bi-LSTM/ конкатенация/ поздний/ FC classifier	SEED-IV, DEAP	ЭЭГ и взгляд	SEED-IV: 4 кат. DEAP: ПБВ*	accuracy: SEED-IV: 93.9 % DEAP: 83.5 %	[70], 2017
CNN, BERT, RNN/ attention/ поздний/ FC classifier	Aff-Wild2	видео и звук, распознанный текст	ПБВ	$\rho_c = 0.619$	[73], 2024
ПБВ — пространство эмоций возбуждения и валентности ПБВ* — ПБВ упрощенное до двух задач бинарной классификации ρ_c — коэффициент корреляции конкордации					

Средний по возбуждению и валентности коэффициент корреляции конкордации продемонстрированный в данной работе равен 0.619. Интересным решением в данной работе является применение распознавания речи (Vosk toolkit), т. к. в используемом в данной работе наборе данных (Aff-Wild2) текстовая модальность не представлена. Исследователям удалось продемонстрировать улучшение качества классификации при добавлении текстовой модальности данных путём распознавания голоса.

Обобщение результатов для рассмотренных алгоритмов приведено в таблице 6.

3. Обсуждение

Методы распознавания эмоций на основе машинного обучения с использованием мультимодальных данных позволяют улучшить качество распознавания эмоций, в сравнении с методами, использующими одну модальность [60, 61, 71]. Такие методы позволяют сделать систему распознавания эмоций более устойчивой к шуму и погрешностям, возникающим при использовании отдельных модальностей.

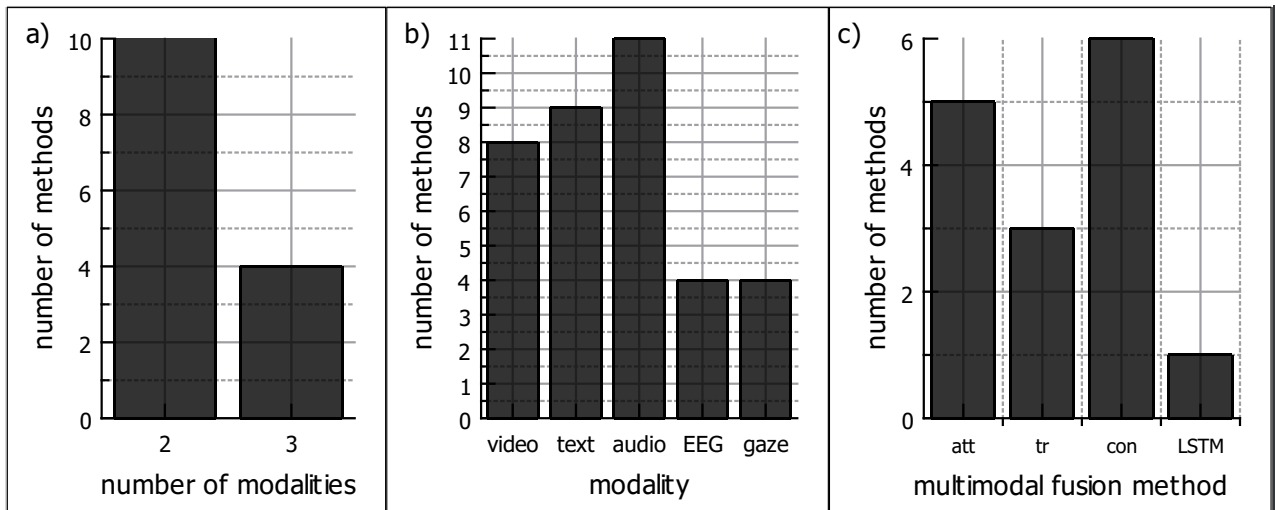


Fig. 10. Comparison of the considered methods for multimodal emotion classification. a — number of modalities used. b — type of modalities used. c — modality fusion method used (att — method based on attention layer, tr — method based on transformer architecture, con — concatenation, LSTM — MMResLSTM).

Рис. 10. Сравнение рассмотренных методов мультимодальной классификации эмоций. а — количество используемых модальностей. б — тип используемых модальностей. с — применяемый метод слияния модальностей (att — метод на основе слоя внимания, tr — метод на основе архитектуры трансформера, con — конкатенация, LSTM — MMResLSTM).

Большинство из рассмотренных методов используют только 2 модальности данных, и лишь некоторые из них используют 3 модальности (рис. 10a). Таким образом, по-прежнему актуальным является вопрос об эффективности применения и методах слияния большего числа модальностей. В ряде работ (см., например, [62]) продемонстрировано улучшение качества классификации при добавлении модальностей, однако вопрос о достаточном и избыточном числе используемых модальностей остаётся открытым.

Для мультимодальной классификации эмоций чаще всего применяются такие модальности, как звук, текст и видео (рис. 10b). Это, в первую очередь, обусловлено простотой получения данных этих модальностей в реальных условиях и наличием эффективных архитектур нейронных сетей для выделения признаков. Однако в работах [61, 68, 70] продемонстрировано достижение схожего или лучшего качества классификации при использовании модальностей физиологических данных (ЭЭГ, взгляд), в сравнении с методами рассмотренными выше. С учётом распространения носимых устройств, получающих физиологические данные пользователей в непрерывном режиме в реальных условиях, методы мультимодальной классификации, использующие текст, звук и видео совместно с физиологическими данными, являются чрезвычайно перспективными для изучения и реализации.

Большинство методов рассмотренных в данной статье используют для слияния модальностей данных конкатенацию признаков полученных из данных разных модальностей (рис. 10c), при этом методы использующие более сложные способы слияния модальностей [56, 69] демонстрируют лучшее качество классификации. Разработка эффективного метода слияния модальностей для классификации эмоций остаётся перспективным направлением исследований.

Помимо методов слияния модальностей важным направлением исследований является разработка методов обучения мультимодальных нейронных сетей для классификации эмоциональных состояний. Как показано в работах [63, 64], добавление модальностей данных может повлечь за собой ухудшение качества классификации в сравнении с одномодальными методами, вызванное конку-

ренцией элементов нейронной сети, отвечающих за различные модальности в процессе обучения. Одним из возможных решений этой проблемы может являться заморозка весов предобученных нейронных сетей, использующихся для выделения признаков из данных отдельных модальностей (см., например, [9, 72]), но такой подход накладывает ограничение на структуру мультимодальной нейронной сети. Существуют специализированные методы, балансирующие обучение нейронной сети для нескольких модальностей [65, 66]. Разработка подобных методов является важным направлением исследований для дальнейшего развития мультимодального подхода в классификации эмоциональных состояний.

На данный момент существует значительное количество открытых наборов мультимодальных данных эмоциональных состояний. При этом, данные наборы зачастую содержат данные малого числа испытуемых (10–30) [24, 27, 30, 31], что поднимает вопрос о переносимости моделей машинного обучения, обученных на этих наборах данных, на новых испытуемых. Наборы данных, содержащие данные большого числа испытуемых (1000) (например, [32]), как правило, содержат малое число модальностей (видео, звук и текст) и достаточно простую парадигму эксперимента (монолог). Таким образом, для активного развития направления классификации эмоций с использованием мультимодальных данных существует потребность в создании объемных наборов данных, содержащих большое число модальностей, и доступных для получения в реальных условиях.

Заключение

Таким образом, использование мультимодальных данных для распознавания эмоций с помощью нейронных сетей является активно развивающимся и перспективным направлением научных исследований.

В области разработки нейросетевых архитектур для классификации эмоциональных состояний на основе мультимодальных данных достигнут значительный прогресс, при этом остаются активно развивающиеся направления:

- разработка методов обучения мультимодальных нейронных сетей;
- разработка эффективных методов слияния модальностей;
- изучение совместимости различных модальностей данных, достаточного и избыточного числа применяемых модальностей;
- изучение проблемы применения моделей, обученных на доступных на данный момент данных, в реальных условиях;
- выделение наиболее перспективного для применения в реальных условиях подхода к классификации эмоциональных состояний: категориальный или в ПВВ.

На данный момент существует значительное число открытых наборов мультимодальных данных эмоциональных состояний, позволяющих разрабатывать новые мультимодальные методы машинного обучения. Однако важным направлением для дальнейшего развития методов мультимодальной классификации эмоций является создание больших наборов данных, полученных в условиях, максимально приближенных к реальным и содержащих не только видео и звуковые модальности, но и физиологические данные, полученные, в том числе, с помощью носимых устройств.

References

- [1] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, “Emotion recognition using facial expressions”, *Procedia Computer Science*, vol. 108, pp. 1175–1184, 2017.
- [2] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, “Emotion recognition by speech signals”, in *Proceedings of the 8th European Conference on Speech Communication and Technology*, 2003, pp. 125–128.

- [3] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, and S. Zeebaree, "Multimodal emotion recognition using deep learning", *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 73–79, 2021.
- [4] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (provably)", *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 944–10 956, 2021.
- [5] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis", *Journal on Multimodal User Interfaces*, vol. 3, pp. 33–48, 2010.
- [6] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures", in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2016, pp. 1–9.
- [7] H. Huang, Z. Hu, W. Wang, and M. Wu, "Multimodal emotion recognition based on ensemble convolutional neural network", *IEEE Access*, vol. 8, pp. 3265–3271, 2019.
- [8] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 6, pp. 112–121, 2021.
- [9] H.-D. Le, G.-S. Lee, S.-H. Kim, S. Kim, and H.-J. Yang, "Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning", *IEEE Access*, vol. 11, pp. 14 742–14 751, 2023.
- [10] H. Al Osman and T. H. Falk, "Multimodal affect recognition: Current approaches and challenges", in *Emotion and Attention Recognition Based on Biological Signals and Images*. 2017, pp. 59–86.
- [11] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Attention driven fusion for multi-modal emotion recognition", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 3227–3231.
- [12] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances", *IEEE access*, vol. 7, pp. 100 943–100 953, 2019.
- [13] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors", *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2016.
- [14] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schüller, "Enhanced semi-supervised learning for multimodal emotion recognition", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5185–5189.
- [15] A. Althnian *et al.*, "Impact of dataset size on classification performance: An empirical evaluation in the medical domain", *Applied sciences*, vol. 11, no. 2, p. 796, 2021.
- [16] Z. Zhang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3438–3446.
- [17] V. Ovsyannikova, "K voprosu o klassifikacii emocij: Kategorial'nyj i mnogomernyj podhody", *Finansovaya Analitika: Problemy i Resheniya*, no. 37, pp. 43–48, 2013, in Russian.
- [18] A. Ortony and T. J. Turner, "What's basic about basic emotions?", *Psychological review*, vol. 97, no. 3, pp. 315–331, 1990.
- [19] P. Ekman, "An argument for basic emotions", *Cognition & emotion*, vol. 6, no. 3, pp. 169–200, 1992.
- [20] C. E. Izard, *The psychology of emotions*. Springer Science & Business Media, 1991.

- [21] W. Wundt, “Outlines of psychology”, in *Wilhelm Wundt and the Making of a Scientific Psychology*. 1980, pp. 179–195.
- [22] J. A. Russell, “Core affect and the psychological construction of emotion.”, *Psychological review*, vol. 110, no. 1, p. 145, 2003.
- [23] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, “Amigos: A dataset for affect, personality and mood research on individuals and groups”, *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 479–493, 2018.
- [24] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, “Emotionmeter: A multimodal framework for recognizing human emotions”, *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2018.
- [25] V. Markova, T. Ganchev, and K. Kalinkov, “Clas: A database for cognitive load, affect and stress recognition”, in *Proceedings of the International Conference on Biomedical Innovations and Applications*, 2019, pp. 1–4.
- [26] K. Sharma, C. Castellini, E. L. Van Den Broek, A. Albu-Schaeffer, and F. Schwenker, “A dataset of continuous affect annotations and physiological signals for emotion analysis”, *Scientific data*, vol. 6, no. 1, p. 196, 2019.
- [27] S. Koelstra *et al.*, “Deap: A database for emotion analysis; using physiological signals”, *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.
- [28] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, “DECAF: MEG-based multimodal database for decoding affective physiological responses”, *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 209–222, 2015.
- [29] C. Y. Park *et al.*, “K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations”, *Scientific Data*, vol. 7, no. 1, p. 293, 2020.
- [30] C. Busso *et al.*, “IEMOCAP: Interactive emotional dyadic motion capture database”, *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.
- [31] S. Katsigiannis and N. Ramzan, “DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices”, *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2017.
- [32] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph”, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.
- [33] D. Kollias *et al.*, “The 6th affective behavior analysis in-the-wild (abaw) competition”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4587–4598.
- [34] T. Bänziger, D. Grandjean, and K. R. Scherer, “Emotion recognition from expressions in face, voice, and body: The multimodal emotion recognition test (MERT)”, *Emotion*, vol. 9, no. 5, p. 691, 2009.
- [35] R. M. Sabour, Y. Benezeth, P. De Oliveira, J. Chappe, and F. Yang, “UBFC-Phys: A multimodal database for psychophysiological studies of social stress”, *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 622–636, 2021.
- [36] J. F. Cohn, Z. Ambadar, and P. Ekman, “Observer-based measurement of facial expression with the facial action coding system”, in *The Handbook of Emotion Elicitation and Assessment*. 2007, pp. 203–221.

- [37] S. Gupta, “Facial emotion recognition in real-time and static images”, in *Proceedings of the 2nd International Conference on Inventive Systems and Control*, 2018, pp. 553–560.
- [38] A. R. Khan, “Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges”, *Information*, vol. 13, no. 6, p. 268, 2022.
- [39] A. Savran *et al.*, “Bosphorus database for 3D face analysis”, in *Biometrics and Identity Management: First European Workshop, BIOID 2008, Roskilde, Denmark, May 7-9, 2008. Revised Selected Papers 1*, 2008, pp. 47–56.
- [40] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, “Speech based human emotion recognition using MFCC”, in *Proceedings of the International Conference on Wireless Communications, Signal Processing and Networking*, 2017, pp. 2257–2260.
- [41] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cleder, “Automatic speech emotion recognition using machine learning”, in *Social Media and Machine Learning [Working Title]*, 2019, hal–02 432 557.
- [42] A. Tripathy, A. Agrawal, and S. K. Rath, “Classification of sentiment reviews using n-gram machine learning approach”, *Expert Systems with Applications*, vol. 57, pp. 117–126, 2016.
- [43] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, “Transformer models for text-based emotion detection: A review of BERT-based approaches”, *Artificial Intelligence Review*, vol. 54, no. 8, pp. 5789–5829, 2021.
- [44] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, *MELD: A multimodal multi-party dataset for emotion recognition in conversations*, 2019. arXiv: [1810.02508 \[cs.CL\]](https://arxiv.org/abs/1810.02508).
- [45] M. Egger, M. Ley, and S. Hanke, “Emotion recognition from physiological signal analysis: A review”, *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35–55, 2019.
- [46] H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, “A transformer-based model with self-distillation for multimodal emotion recognition in conversations”, *IEEE Transactions on Multimedia*, vol. 26, pp. 776–788, 2023.
- [47] S. Anand, N. K. Devulapally, S. D. Bhattacharjee, and J. Yuan, “Multi-label emotion analysis in conversation via multimodal knowledge distillation”, in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6090–6100.
- [48] Y. Li *et al.*, “A novel bi-hemispheric discrepancy model for EEG emotion recognition”, *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 2, pp. 354–367, 2020.
- [49] B. Fu, C. Gu, M. Fu, Y. Xia, and Y. Liu, “A novel feature fusion network for multimodal emotion recognition from EEG and eye movement signals”, *Frontiers in Neuroscience*, vol. 17, p. 1 234 162, 2023.
- [50] C. Tan, M. Šarlija, and N. Kasabov, “NeuroSense: Short-term emotion recognition and understanding based on spiking neural network modelling of spatio-temporal EEG patterns”, *Neurocomputing*, vol. 434, pp. 137–148, 2021.
- [51] P. Bhattacharya, R. K. Gupta, and Y. Yang, “Exploring the contextual factors affecting multimodal emotion recognition in videos”, *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1547–1557, 2021.
- [52] H. Zhang, “Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder”, *IEEE Access*, vol. 8, pp. 164 130–164 143, 2020.

- [53] F. Al-Naima, S. Y. Ameen, and A. F. Al-Saad, “Destroying steganography content in image files”, in *Proceedings of IEEE Fifth International Symposium on Communication Systems, Networks and Digital Signal Processing*, 2006, pp. 1–4.
- [54] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, “Multi-sequence LSTM-RNN deep learning and metaheuristics for electric load forecasting”, *Energies*, vol. 13, no. 2, p. 391, 2020.
- [55] T. D. Nguyen, “Multimodal emotion recognition using deep learning techniques”, Ph.D. dissertation, Queensland University of Technology, 2020.
- [56] A. Joshi, A. Bhat, A. Jain, A. V. Singh, and A. Modi, *COGMEN: COntextualized GNN based multimodal emotion recognition*, 2022. arXiv: [2205.02455 \[cs.CL\]](#).
- [57] J.-H. Lee, H.-J. Kim, and Y.-G. Cheong, “A multi-modal approach for emotion recognition of TV drama characters using image and text”, in *Proceedings of the IEEE International Conference on Big Data and Smart Computing*, IEEE, 2020, pp. 420–424.
- [58] W. Ai, F. Zhang, Y. Shou, T. Meng, H. Chen, and K. Li, “Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum”, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 11 418–11 426.
- [59] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, *Jointly fine-tuning “BERT-like” self supervised models to improve multimodal speech emotion recognition*, 2020. arXiv: [2008.06682 \[eess.AS\]](#).
- [60] D. Krishna and A. Patil, “Multimodal emotion recognition using cross-modal attention and 1D convolutional neural networks”, in *Proceedings of the Interspeech*, 2020, pp. 4243–4247.
- [61] B. Fu, C. Gu, M. Fu, Y. Xia, and Y. Liu, “A novel feature fusion network for multimodal emotion recognition from EEG and eye movement signals”, *Frontiers in Neuroscience*, vol. 17, p. 1 234 162, 2023.
- [62] Z. Pan, Z. Luo, J. Yang, and H. Li, *Multi-modal attention for speech emotion recognition*, 2020. arXiv: [2009.04107 \[eess.AS\]](#).
- [63] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, “Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably)”, in *International Conference on Machine Learning*, 2022, pp. 9226–9259.
- [64] W. Wang, D. Tran, and M. Feiszli, “What makes training multi-modal classification networks hard?”, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 695–12 705.
- [65] H. Li, X. Li, P. Hu, Y. Lei, C. Li, and Y. Zhou, “Boosting multi-modal model performance with adaptive gradient modulation”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 214–22 224.
- [66] K. Kontras, C. Chatzichristos, M. Blaschko, and M. De Vos, *Improving multimodal learning with multi-loss gradient modulation*, 2024. arXiv: [2405.07930 \[cs.MM\]](#).
- [67] A. Baeviski, Y. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations”, *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [68] W. Liu, W.-L. Zheng, and B.-L. Lu, “Emotion recognition using multimodal deep learning”, in *Proceedings of the Neural Information Processing: 23rd International Conference, Part II* 23, 2016, pp. 521–529.
- [69] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, “Emotion recognition using multimodal residual LSTM network”, in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 176–183.

- [70] H. Tang, W. Liu, W.-L. Zheng, and B.-L. Lu, “Multimodal emotion recognition using deep neural networks”, in *Proceedings of the Neural Information Processing: 24th International Conference, Part IV 24*, 2017, pp. 811–819.
- [71] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, “End-to-end multimodal emotion recognition using deep neural networks”, *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.
- [72] W. Zhang, B. Ma, F. Qiu, and Y. Ding, “Multi-modal facial affective analysis based on masked autoencoder”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5793–5802.
- [73] R. G. Praveen and J. Alam, “Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4803–4813.