Санкт-Петербургский государственный университет
**Факультет прикладной математики — процессов управления**
**Кафедра моделирования экономических систем**

## Зырянов Дмитрий Сергеевчич

**Выпускная квалификационная работа бакалавра**

# Распознавание лицевой экспрессии с помощью нейронных сетей

Направление 010300
Фундаментальная информатика и информационные технологии

Заведующий кафедрой,
доктор физ.-мат. наук,
профессор — Андрианов Сергей Николаевич

Научный руководитель,
доктор физ.-мат. наук,
профессор — Андрианов Сергей Николаевич

Рецензент,
Старший преподаватель кафедры
технологии программирования — Севрюков Сергей Юрьевич

Санкт-Петербург
2017

# Содержание

# Введение

Что такое эмоции? Сможем ли мы верно ответить на этот вопрос? А какова вероятность, что мы точно сможем распознать лицевую экспрессию у своего собеседника?

Данный вопрос давно интересует не только психологов современного мира, так же этой проблемой давно занимаются и исследуют ученые всего мира. Проведенные исследования на эту тему были затронуты еще в книге «Выражение эмоций у человека и животного» Чарльза Дарвина [1], где он утверждает, что эмоции скорее варьируются от вида к виду, то есть это видовой признак, чем культурный. В начале 20 века существовала достаточно фундаментальная теория о том, что не существует «универсальных» эмоций. Данная теория опиралась на зависимость представителя от типа культуры. На эту же тему в 1969 году была издана одна из самых важных литературных работ, проделанная психологами и учеными. Это исследование Пола Экмана и Уоллеса Фризена [2] [3], которые были явными оппонентами существующих канонов эмоционального происхождения. Они доказали, что существует 6 универсальных эмоций, которые не зависят от культурной и национальной принадлежности: удовольствие, страдание, страх, гнев, удивление и отвращение. Конечно же, к данным эмоциям всегда относят нейтральное состояние лица человека.

В последние годы распознавание эмоций стало принимать популярный характер не только в методах и специфике психологии и социологии. С учетом всемирной и всепоглощающей идеи об искусственном интеллекте ученые и программисты всего мира пришли к консенсусу необходимости достаточно точного определения эмоций с помощью программных средств. Распознавание любого вида проявления человеческого отклика и реакции –это одно из самых важных препятствий для создания искусственного интеллекта. На данный момент такой вектор исследования очень важен и имеет огромное значение для всего человечества, особенно в отраслях бизнеса и военных структурах.

Принимая во внимание нынешнее пристрастие к автоматизации процессов, ученые, и не только, стали заниматься проблемами распознавания и детектирования большого количества объектов, окружающих нас. И, в какой-то мере, это приносит хорошие результаты. Распознавание номеров у машин уже вошло в обиход и активно используется правоохранительными органами, детектирование лиц с помощью программ и компьютера — уже давно не новинка, так как во многих крупных компаниях стоит система безопасного входа по определению сотрудника посредством фотографий. Но один из самых важных прорывов во взаимодействии безопасности и детектирования был совершен при введении на особо важные объекты камер с распознаванием отрицательных, враждебно настроенных эмоций, которые помогут вычислить преступника, грабителя или террориста. Нельзя не упомянуть про прорывы в распознавании голоса, тональности текста, речи и так далее. Исходя из этого, видно, что данное направление не стоит на месте и успешно развивается. На данный момент значительную нишу в этом научном направлении занимает машинное обучение. Уже давно признали, что машинное обучение одно из самых успешных, конструктивных, универсальных и оптимальных методов для распознавания объектов на фотографии, видео фрагментах, фильмах и электронных записях. Машинное обучение показало и продолжает показывать огромное увеличение точности алгоритмов по детектированию. Но главенствующее положение на данный момент в области машинного обучения занимают нейронные сети, которые безупречно доказали свою эффективность, и не перестают улучшать результат, поэтому мы будем использовать именно этот метод. Данный метод в основном требует либо огромного количества данных, которые послужат тренировочными и тестовыми образцами, либо идеально выстроенной нейронной сети, которая учитывает все нюансы и специфические моменты требуемой задачи. Важной проблемой в области распознавания лицевых экспрессий—это отсутствие согласованной и полностью учитывающей всех особенностей эталонной базы данных, на которой бы строились все алгоритмы. Данный эталон служил бы некоторым главен-

ствующим инструментом, который обеспечивал бы возможность сравнительной характеристики ныне существующих методов и подходов. Было сделано множество попыток создать данную базу фотографий, но ученые и программисты так и не пришли к общему паттерну. Возможно, проблема кроется в глобализации маркетинга и поискам выгоды в каждой научной отрасли.

Идея нейронных сетей была взята из физиологии, ведь, как всем известно, прототип этой архитектуры был получен из примера функционирования человеческого мозга. Человеческая биологическая нейронная сеть была взята за основу. Ученые пытались смоделировать работу нашего мозга и протекающих в нем реакций, поэтому данный метод и нашел применение в основном в распознавании объектов, образов и в задачах прогнозирования. Существует несколько методов, взятых из научной литературы, для распознавания эмоций на лице человека:

1. Статичный метод В данном методе классификатор, который используют авторы [4], распознает каждый кадр и сравнивает его с заявленными вначале эмоциями, и на основе каждого кадра делает вывод. Обычно в таких подходах используется Наивный Байесовский классификатор [4]. В то же время, он использует строгое и нереалистичное предположение о том, что все признаки не взаимосвязаны друг с другом и являются новым классом.

2. Динамический метод В этом методе, классификаторы принимают во внимание временный шаблон, полученный до этого, для отображения выражения лица. Скрытая Марковская модель [5] наилучше подходит для такой реализации. СММ отслеживает сгенерированные до этого состояния, и на основе этого достаточно оптимально выводит полученный результат. На основе данного алгоритма в работе получилось даже отследить долгий фрагмент и получить необходимый прогноз по отслеживанию эмоции.

Работы по теме распознавания можно обычно разделить на два типа:

- Сравнение методов по обработке и определение эмоции

- Введение нового метода, полученного на основе предыдущих работ, который усовершенствует предыдущие результаты.

Иногда в последнее десятилетие можно выделить еще один тип, который описывает разницу в осуществлении разных подходах на разном типе машин и разном типе аппаратурной части (обучение на CPU и GPU [6]).

## Приоритетные задачи

Задача идентификации эмоций обычно решается в два этапа, каждый из которых бесспорно важен и определяет успешность результата, первый — распознавание лица на изображении, второй — распознавание лицевой экспрессии на основе полученного лица. В своей работе я рассмотрю метод, который будет распознавать эмоцию человека из полученного лица при помощью построенной и обученной на выборке нейронной сети.

Мой приоритет — улучшить значение точности алгоритмов по сравнению с человеческой точностью на выборке, то есть достигнуть того, что компьютер решает проблему распознавания не хуже человеческого глаза. Самая основополагающая задача–сгенерировать подход, который одинаково будет распознавать эмоции, вне зависимости от национальности и возраста. Дополнительные задачи, которые необходимо выполнить:

- Необходимо изучить и исследовать нейронные сети в области детектирования экспрессий, и вообще в области распознавания
- Провести качественный обзор и сравнение уже существующих методов, с помощью которых уже была решена поставленная мною задача
- Выделить проблемы данной научной области и обозначить перспективы развития

Для дальнейшего прорыва в получении и создании искусственного интеллекта необходимо понимание особенностей работы мозга, в том числе распознавания объектов. Поэтому вектор данной работы очень востребован и популярен.

# 1. Формализация задачи

## 1.1 Постановка задачи

Данная задача была уже решена многими крупными компаниями, которые занимаются разработками систем по детектированию и распознаванию. Поэтому, в первую очередь данная работа носит в себе исследовательский характер. Поставлена цель создать алгоритм, изучить методы и средства разработки по идентификации эмоций. Конечный алгоритм, как и любой другой алгоритм по распознаванию лицевых экспрессий, будет разделен на три основных последовательных этапа:

1. Определение лица

   Данный этап в работе требует тщательной работы, потому что только эта одна тема рассматривается как отдельная серьезная работа, над решением которой уже давно работают ученые. Это задача требует учета многих особенностей. Например, присутствует ли лицо на изображении или полученном кадре? Если был получен положительный ответ, то с помощью программных и аппаратных средств должен быть получен размер лица и его положение (то есть, его координаты в двухмерной плоскости) на изображении. Так же сюда может быть включен подпункт про отслеживание лица, то есть изменение его положения в плоскости по времени.

   Однако, задача построения качественного детектора не будет решаться в работе.

2. Определение и описание особенностей лица человека

   На данном этапе работы подход должен определять положение в плоскости значимых частей лица, которые будут задействованы в дальнейшем распознавании эмоций и лицевых экспрессий. В число этих важных частей лица могут входить: рот, глаза, брови, нос и так далее. Варьируются эти варианты в зависимости с выбранном или созданном алгоритме при определении. Для решения этой части необходимо

построение архитектуры нейронной сети, которая будет генерировать матрицу весов для обучающей выборки, учитываю все эти особенности. Нейронная сеть–это некий «черный ящик», то есть пользователь не имеет возможности узнать, какие ключевые элементы в основном учитывает построенная сеть.

3. Распознавание эмоции

   Эта часть работы уже финальная и требует работы с полученными данными. На основе матрицы весов, полученной с помощью обучения сети, алгоритм должен сделать прогноз с помощью нейронной сети. Согласно информации, данной нам с первых двух частях работы, наш алгоритм должен выполнять свою первоначальную и незыблемую функцию–определить эмоцию на изображение.

## 1.2 Обзор существующих алгоритмов

Анализ распознавания лицевых экспрессий, или, как тривиальнее, эмоций, является сложной и комплексной задачей, поэтому существует огромное количество методов и алгоритмов, рассмотренных в работах разных авторов. Некоторые методы очень спорны и имеют существенные недостатки, так как авторы пытаются найти все более изощренный метод, который докажет свою состоятельность и выявит недостатки предыдущих алгоритмов. Есть некоторые особенности лица человека, из–за которых не существует на данный момент эталонного способа решения задачи. Часто многие методы не учитывают аксессуаров и лицевой растительности, типа бороды, усов, неравномерных бровей и объемных причесок, закрывающих или мешающих распознаванию. Плюс к тому, эмоция очень абстрактное выражение, которое очень изменяется от человека к человеку. Часто мы сталкиваемся с проблемой непонимания некоторой экспрессии, например, открытый рот не всегда означает удивление. Однако, основной недостаток, как было сказано выше—отсутствие эталонной обучающей выборки, на основе которой создается алгоритм. Методы, расписанные ниже, чаще применяются к детектированию лиц на изображении, но многие были при-

менены и к эмоциональному распознаванию.

1. Распознавание эмоций в реальном времени на основе нечёткой логики
   Данный алгоритм учитывает особенности нечеткой логики (Fuzzy Logic) [7]. Авторы данного алгоритма показывают преимущества решения подобных плохо формализуемых задач, таких как определение эмоций. Методы адаптивного управления строят самообучающиеся распознающе–управляющие комплексы, которые одновременно решают множество задач, таких как:

   - автоматическая классификация
   - распознавание образов на изображении
   - качественная оценка полученных значений (в данном случае определение эмоции)
   - небольшая ошибка прогнозирования
   - принятие окончательного решения

   Данный метод доказал свою состоятельность и показал хорошие погрешности на тестируемых данных, которые привели ученых к выводу, что данный подход необходимо изучать и развивать. Из негативных сторон можно выделить то, что все эти задачи не всегда целесообразно и выгодно решать в одной системе. Иногда проще и конструктивнее создание комплексной системы, учитывающей все особенности. На Рис. 1 представлен пример определения эмоции с помощью этого метода.

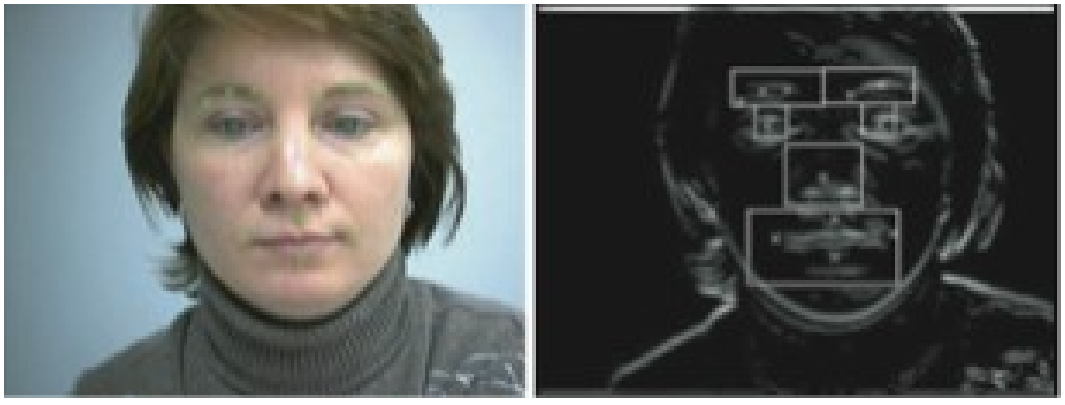

Рис. 1: Модель эмоций на основе Fuzzy Logic

2. Распознавание эмоции по цвету лица человека

   Метод (Emotion Identification Method using RGB information of Human Face [8]), использованный этими исследователями учеными, считается очень спорным и неоднозначным. Однако, он также показал определенно новый и ранее незатронутый подход к изучению данной проблемы. Авторы подходят к эмоциям на лице человека с иной стороны. В этой статье за основу берется цвет или температура полученного лица, которая определяет кровяное давление, то есть используется физиологический аспект в формировании эмоций. Полученные значения разделяется на два подтипа: «теплые» и «холодные». Цвет человеческой кожи определяется из преобладания гемоглобина и меланина. Плюс к тому, вегетативная нервная система делится на симпатическую и парасимпатическую. В основу взят тот факт, что преобладание симпатической нервной системы, как правило, вызывает стресс или гнев, что в данной работе символизируется «холодным» результатом. С другой стороны, с преобладанием парасимпатической нервной системы, люди обычно чувствуют себя спокойно, расслабленно, что в значениях получается «теплым» результатом.На Рис. 2 показана таблица измерения кровяного давления при изменении темературы воздуха.

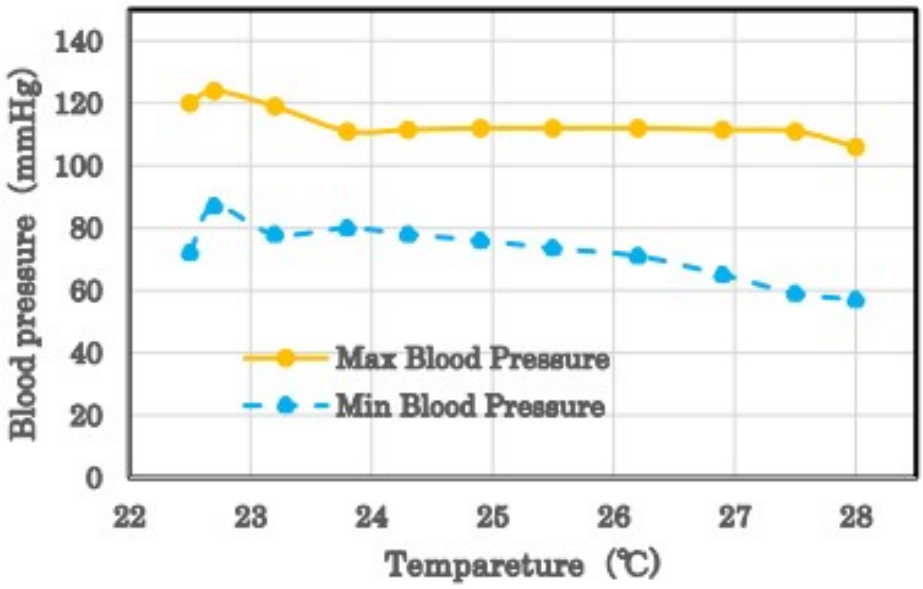


Рис. 2: Пример детектирования кровяного давления

Данный метод сложно охарактеризовать как идеально верный, так как мы можем видеть субъективность самого определения, ведь авторы не учитывают всех особенностей физиологии. Хоть человеческий вид и теплокровный, но у каждого из представителей разный оптимальный температурные график. Организм своеобразно реагирует на внешние признаки, вызываю при этом большую ошибку и погрешность данного метода. Первоначально этот алгоритм был проектом для создания «умных» кондиционеров в зданиях, поэтому и используется данный метод для решения задачи.

3. Метод главных компонент

   Одним из самых популярных и проработанных подходов является метод главных компонент. Метод главных компонент (Principal Component Analysis, PCA) [9] в основном применяется для сжатия или снижения признаков без существенной потери признаков. В основу этого был и положен метод определения лица человека, у которого многие несущественные признаки можно отбросить, то есть провести сжатие. Основной целью данного метода является уменьшение количества признаков таким образом, чтобы оно наилучшим способом имело возможность описать «типичный» эталон. Проделанная работа обычно сводится к тому, что вся наша обучающая выборка (в данном случае лиц) преобразовывается в одну общую матрицу данных, где каждое изображение кодируется в собственной строке. Для успешного выполнения этого алгоритма все изображения должны быть приведены в одному размеру и имели нормированную гистограмму.

   Для каждого изображения путем вычисления получаются главные компоненты. Обычно таких ключевых значений от 5 до 200. Остальные компоненты берутся за мелкие различия лицами иной шум. Далее с помощью какой–либо метрики (обычно Евклидово расстояние) применяется сравнение главных компонент полученного изображения с компонентами других фрагментов.

   Этот метод достаточно оптимальный, но для его идеальной реализации

необходимо, чтобы все входные данные были с одинаковой освещенностью, иначе возможно появление дополнительных ненужных компонент, которые могут затруднить определение лицевой экспрессии. Плюс к тому, данный подход очень чувствителен к появлению аксессуаров и растительности на лице. Но самый главный недостаток–высокая трудоемкость этого подхода. Иногда определение набора собственных векторов очень ресурсоемко и занимает большое количество времени, пример можно увидеть на Рис. 3. Существуют способы, чтобы распознавание и вычисление происходило быстрее.

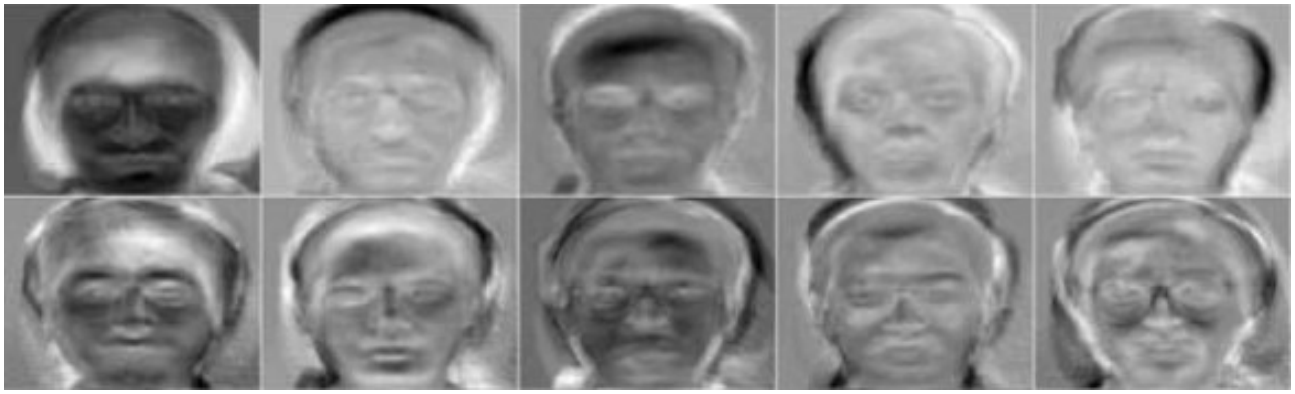

Рис. 3: Пример первых десяти собственных векторов (собственных лиц), полученных на обучаемом наборе лиц

4. Скрытые Марковские модели

Данный подход является мощным вычислительным средством для распознавания речи, в первую очередь. В дальнейшем скрытые Марковские модели [10] [11] (Hidden Markov Model, HMM) были применены для распознавания лиц на изображении. В своей сущности HMM способны учитывать непосредственно пространственно–временные характеристики сигналов. Каждая модель определяется выражением $\lambda = \langle A, B, \pi \rangle$, и представляет собой набор N состояний $S = \{S_1, S_2, \ldots, S_N\}$, между которыми существуют переходы. В момент времени система находится только в одном состоянии. Наиболее популярные Марковские модели первого порядка, где каждое состояние зависит только от предыдущего и больше ни от какого другого. Каждый переход записывается символом и, в дальнейшем, определяется им же, и соответствует физическому сигналу с выхода моделируемой системы. Набор $V = \{v_1, v_2, \ldots, v_m\}$ представляет собой набор выходов. Количество данных символов, по определению, составляет $M$ элементов. $B = \{b_{jk}\}$ матрица вероятности нахождения символа в определенном состоянии. Матрица $A = \left\| a_{ij} \right\|$ определяет вероятность перехода из одного состояния в другое. Последнее число $\pi = \{\pi_i\}$ показывает вероятность начальных состояний. Скрытыми данные модели называется по причине того, что обычно последовательность состояний скрыта от наблюдения и неизвестна в определённый момент времени. Также данный подход был усовершенствован и были введена сеть двумерных Марковских моделей, которое имеет способность моделировать искажения по вертикали и горизонтали одновременно. А для уменьшения вычислительной сложности были придуманы псевдодвумерные HMM (P2D-HMM) [12].На Рис. 4 показана такая модель. Если говорить о недостатках данного подхода, то ученые вывели несколько таковых. Главной проблемой HMM является плохая различающая способность, то есть модель не может адекватно выделять классово типичные признаки, по которым в дальнейшем могла бы распределить все данные на типы.

Поэтому при большой выборке данные могут оказаться незначительно отличимыми.

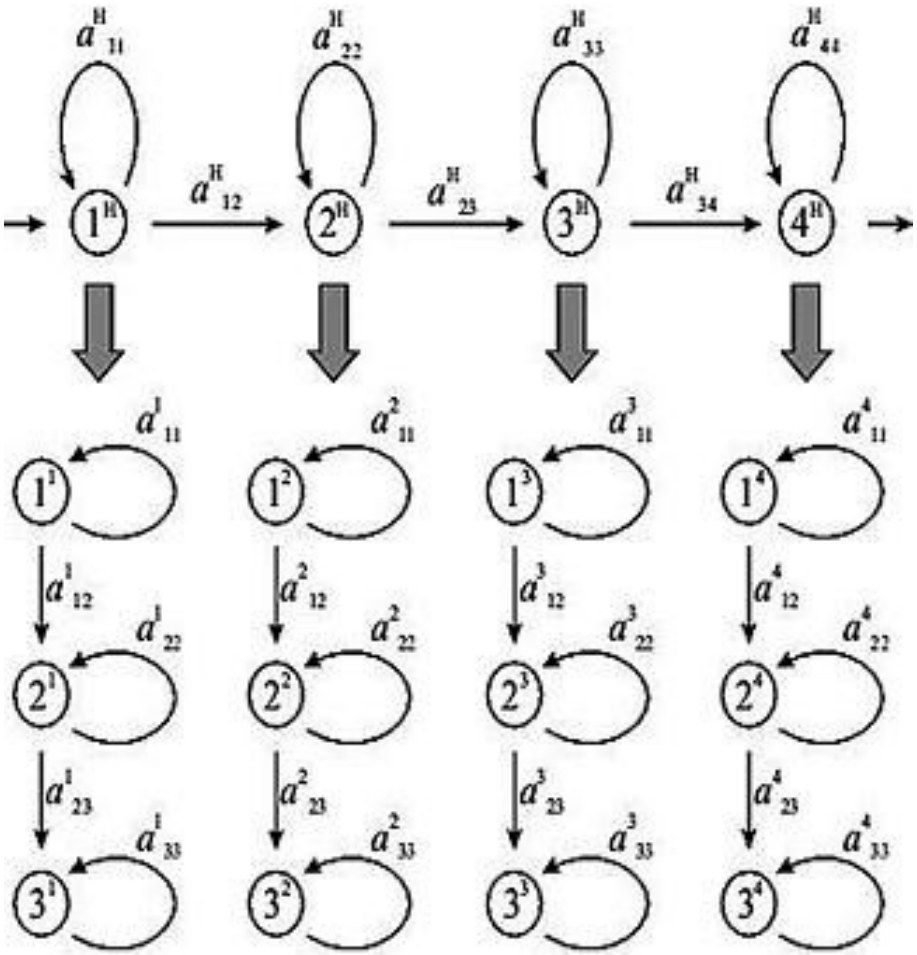

Рис. 4: Псевдодвумерная скрытая Марковская модель

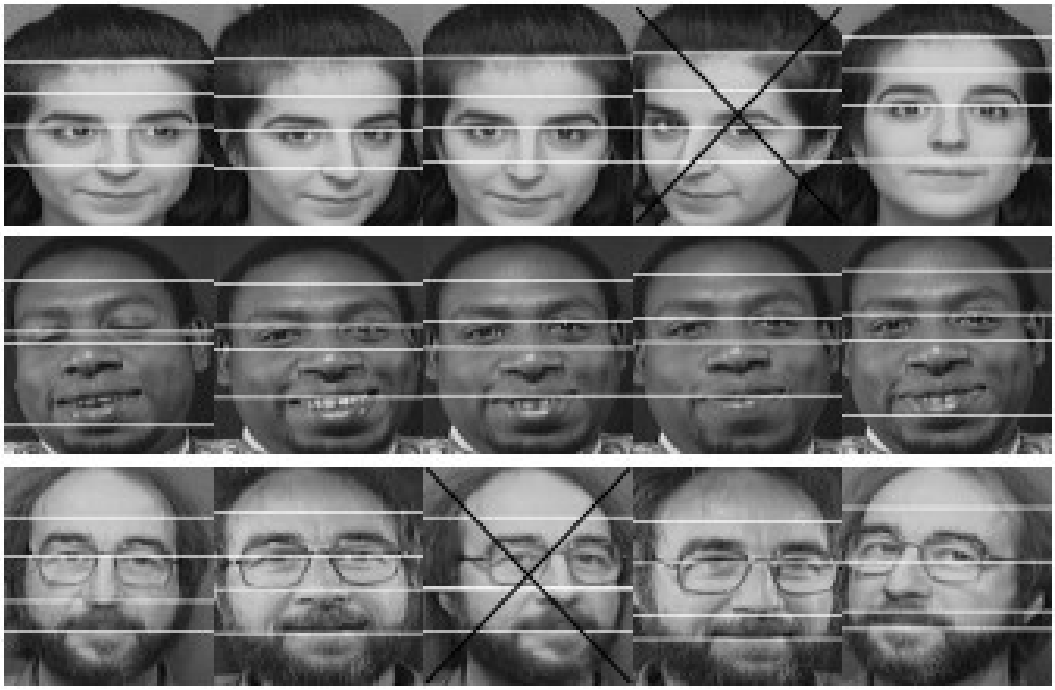

Рис. 5: Пример распознавания лиц с помощью HMM

С одной стороны, существует огромное количество разделения подходов в распознавании эмоции, но обычно всего три неклассических подходов: распознавание уровня серого, движения и частоты. Первый метод использует признак того, что разные эмоции на изображении приводят к разному уровню серого на цифровом фрагменте изображения лица. Движение как характерный признак используется, когда есть возможность отследить изменение положения точек на лице. Если говорить про частоту, то тут диверсия изображений и классифицированная разница между фотографиями одного лица.

Обычно, по определению, принято классифицировать способы распознания на другие фундаментальные различия. Первый тип детектирования–целостные и локальные распознавания. Как мы уже говорили, некоторые методы вычисляют лицо в целом для дальнейшей сравнительной характеристики (Principal Component Analysis PCA [9], Independent Component Analysis ICA, Fisher‘s Linear Discriminants FLD). Локальное детектированием, то есть определение отдельных частей лица (рта, бровей, носа и глаз) и их анализ выполняется методами (Facial Actions Code System FACS, Local PCA, Нейронные сети). Второй тип распозна-

вания определяется выделением изменения лица и движения. Примеры таких подходов: Point Distribution Model(PDM), Active Shape Model(ASM). Методы геометрических характеристик лица основывает третью фундаментальную группу. Данные подходы основаны на получении основных векторов лица, полученных через форму и положения лица человека.

# 2. Теоретические обоснования

## 2.1 OpenCV

Первое, что стоит рассмотреть в этой части, это open–source библиотека OpenCV [13]. Незаменимый инструмент [14] любого начинающего энтузиаста и даже ученого в среде распознавания лиц, образов, эмоций. Данная библиотека в основном направлена на детектировании образов в режиме «real–time», где она показала значительные показатели. Популярность данного инструмента объясняется его бесплатностью, открытостью, кроссплатформенностью и, определенно, большим количеством литературы и туториалов по использованию. Была запущена в 1999 году и до сих пор набирает, и увеличивает свою популярность.

В основном, OpenCV [13] —- библиотека алгоритмов и функций по обработке изображений и видео. Вот некоторые примеры его модулей:

- opencv_video —- занимается анализом и отслеживанием объектов на видео, также устраняет фон
- opencv_gpu —- очень важный модель, который подключает вычисления на GPU за счет CUDA(NVidia)
- opencv_imgproc —- обработка изображений (некоторые фильтры, преобразования и так далее)
- opencv_ml —- инструменты машинного обучения (SVM, дерево принятия решений)

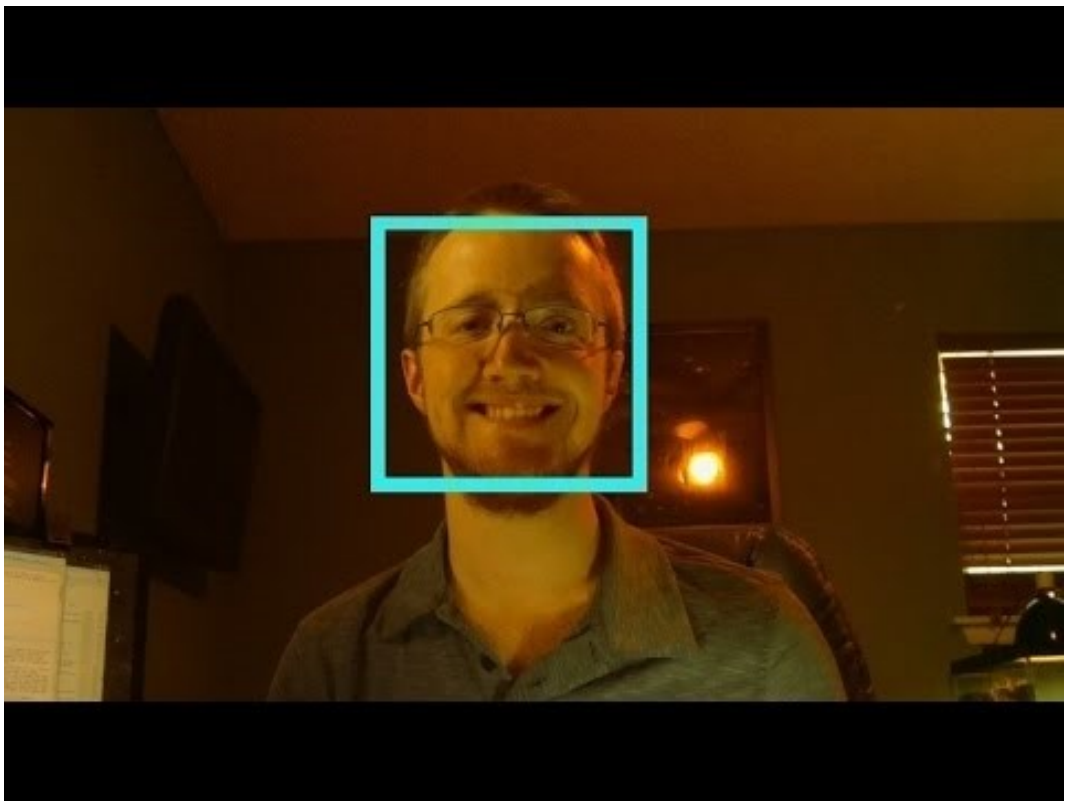

Рис. 6: Пример детектирования лица с помощью OpenCV

Данный метод [14] активно пользуется разработчиками и молодыми учеными в сфере распознавания лиц в последние годы. Уже редко встретишь работы и статьи по этой теме без использования или упоминания каскадов Хаара [15]. Данный инструмент является набором примитивов, которые работают с интенсивностью изображения. В общем случае, это набор смежных прямоугольных областей разных уровней, которые вычисляют сумму интенсивностей пикселей в данной области и подсчитывают разницу между суммами областей. В случае каскадов Хаара мы имеем дело с уровнями $-1$ и $+1$. Примитивы могут быть с разным расположением областей и наклоном. При использовании большого количества примитивов, алгоритм более точно классифицирует объект. На Рис. 7 представлен типичный пример каскадов Хаара. Свертка данными прямоугольными областями дает явное структурное представление объекта на фотографии. Более того, на нашем лице можно выделить четкие переходы в интенсивности областей. Например, невооруженным глазом видно, что область глаз темнее, чем верхняя часть щек и лба. А рот будет светлее, чем нос или щеки. Данная особенность дает объяснение, почему каскады

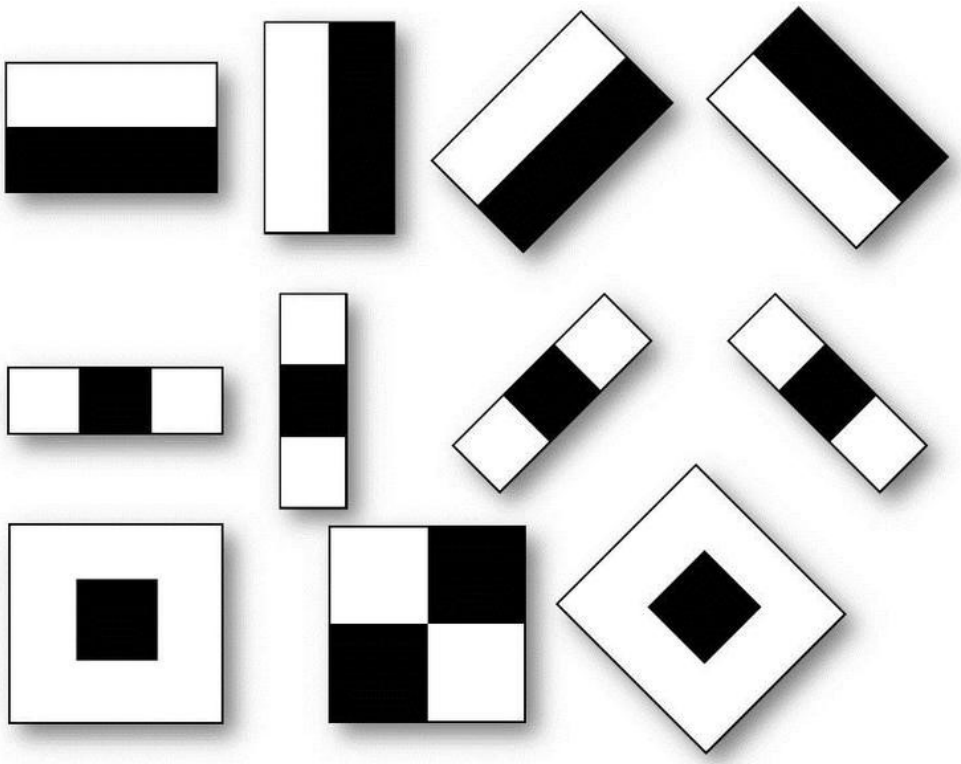

Рис. 7: Пример каскадов Хаара

Хаара нашли применение именно в области распознавания человеческих лиц. Метод Виллы–Джонса [16], который основан на распознавании с помощью прямоугольных примитивов, значительно улучшил скорость и точность распознавания при помощи расчетов с использованием интегрального представления изображений. Интегральное представление изображения — некоторая матрица, по размерности совпадающая с размерностью изображения. Элементы этой матрицы — это сумма интенсивностей пикселей в прямоугольнике с (0,0) до (x,y). И вычисляется по формуле:

$$L(x,y) = \sum_{i=0}^{x}\sum_{j=0}^{y} I(i,j)$$

Преимуществом этого подход является скорость подсчета любой суммы внутри прямоугольника при имеющемся значении смежных прямоугольников по формуле:

$$L(ABCD) = L(A) + L(C) - L(B) - L(D)$$

Подсчет такой суммы показан на Рис. 8 Интегральное представление изоб-

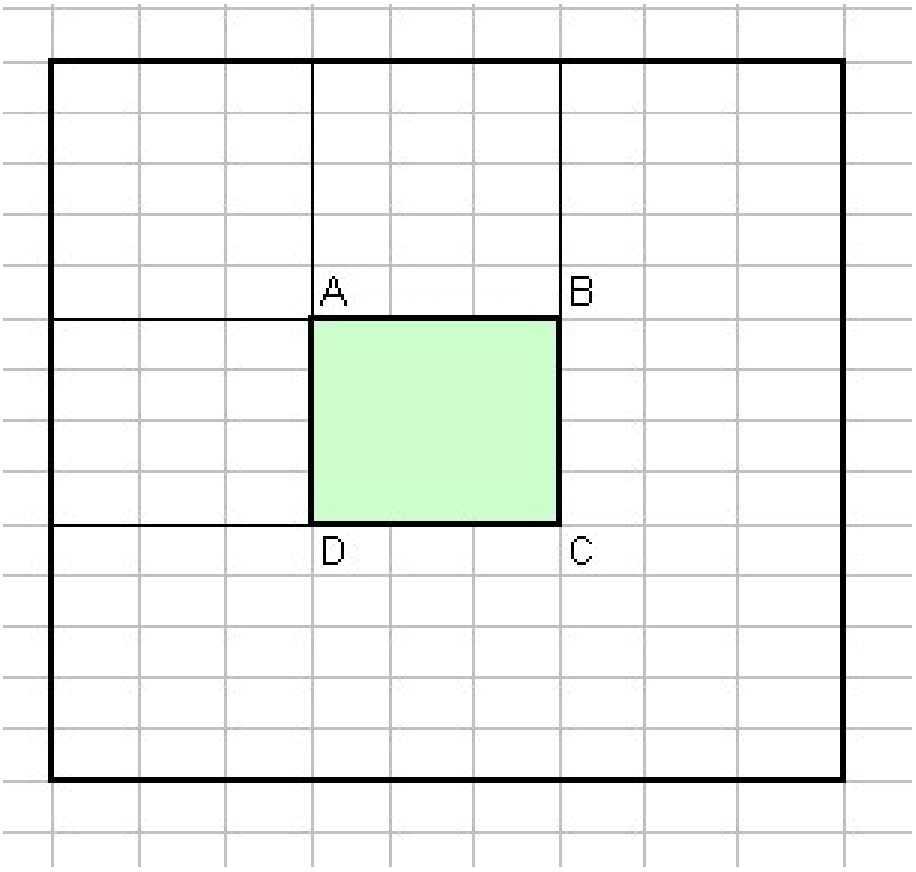


Рис. 8: Подсчет суммы внутри прямоугольника

ражений является основным инструментов в работе с каскадами Хаара и классификаторами, основанными на методе Виола–Джонса. Также эти ученые предложили использовать бустинг AdaBoost [17] для усиления классификаторов при помощи объединения их в один комитет. Недостатки каскадов Хаара:

- Значительная зависимость от положения лица на фотографии
- Невысокая скорость алгоритма при распознавании
- Ухудшение освещения и диверсия фона значительно увеличивают ошибку данного метода

## 2.3 Сверточные нейронные сети

### Нейронные сети в анализе изображений

На данный момент наилучшего результата в методах машинного обучения достигает такая математическая модель нашего мозга как нейронные сети. С развитием аппаратной составляющей, многие крупные компании имеют возможность обучить нейронную сеть на огромных базах данных за небольшой промежуток времени. Данные инновации вывели нейронные сети на первый план, устранив проблему долгого обучения сети. Если заходит разговор об обработке изображений и, в частности, распознавания объектов на фотографиях, то возникает проблема большого объема данных. К примеру, самая распространенная база данных с изображениями содержит несколько миллионов картинок с размерами 256*256 и, записанными в RGB формате. Обычная нейронная сеть будет невероятно долго обучаться даже на супермощных компьютерах и, вдобавок, вероятно, придет к проблеме переобучения, которая заключается в отсутствии возможности воспринимать и классифицировать объекты в общем. Искусственная нейронная сеть на вход получает вектор значений, преобразуя его в дальнейшем через скрытые слои. Эти скрытые слои состоят из пула нейронов, в котором каждый нейрон связан с нейронами предыдущего слоя. Последний слой у нейронной сети — выходной слой. Обычно он представляет собой вектор вероятностей принадлежности к определенному классу. Для примера с картинками с ImageNet в такой полностью связанной архитектуре каждый нейрон будет иметь порядка $256 * 256 * 3 = 196800$f весов. Для полноценной классификации необходимо иметь как минимум несколько таких нейронов. Решением данного казуса является специальная архитектура искусственной нейронной сети — так называемые свёрточные нейронные сети [18] [19]. Данное название прямо происходит от функции, которая выполняется этой сетью — свёрткой. Такой тип сетей относится к разновидности глубинного обучения, использую нелинейные преобразования для получения результата. В сверточных нейронных сетях [18] [19] [20] более разумно организована

архитектура для работы с изображениями. Слои этой нейронной сети расположены в трех измерениях, которые характерны для картинки: высота, ширина и глубина (в отношении изображений это Red, Green, Blue). Нейроны в слое связаны только с некоторой областью в предшествующем слое, а не со всеми как в полностью связанной сети.

### Архитектура CNN

Как было описано выше, у сверточной нейронной сети реализовано несколько слоев, которые преобразуют входные данные в вектор принадлежности к классам. Основные слои, которые используются в сверточной сети — это слой свертки, слой линейной ректификации, слой субдискретизации и полносвязный слой. У каждого из этих своя функции и задача, но в итоге мы получаем сложно структурированную сеть с наилучшими результатами для классификации объектов на изображении. На Рис.9 подробно иллюстрирована архитектура CNN [18] [19]. Рассмотрим подробнее каждый из слоев:

1. Входной

   Первый вход у любого типа нейронной сети — входной. Этот слой хранит необработанные значения пикселей входного изображения. В случае с изображением это некоторое количество значений по ширине, высоте и глубине.

2. Слой свертки

   Сверточный слой — это основной слой в CNN [18] [19]. Данный слой зиждется на операции свертки, которая заключается в том, что по всем входным данным «пробегается» матрица весов ограниченного и небольшого размера, создающая новое значение для некоторого числа пикселей. Такая матрица называется ядром свертки, ее особенностью является то, что весовые коэффициенты такой матрицы определяются во время обучения и заранее неизвестны. К необходимости использования этого слоя можно отнести то, что он значительно уменьшает количество весов, созданных для нейронов, в отличии от полносвязной

нейронной сети.

3. Слой с блоком линейной ректификации(ReLu)

   Данный слой содержит функцию активацию после операции свертки. В сверточной нейронной сети используется вместо привычных гиперболической или сигмоидальной функции активации, функции $f(x) = max(0, x)$. Эта активационная показала очень хорошие результаты на задаче классификации изображений. Более того, выбор был сделан на этой функции активации по причине того, что для большого объема данных она позволяет обучаться сети за адекватное время.

4. Слой субдискретизации

   На данной этапе происходит пулинг полученных данных, то есть подвыборка. На этой стадии матрица характеристик, она же матрица со значениями пикселей уплотняется до матрицы меньшего размера. На примере изображений пулинг уменьшается размерность по ширине и высоте. Такое преобразование делается на уникальных и непересекающихся областях, обычно из прямоугольника 2x2 выбирается один пиксель, значение которого максимальное. Тем самым происходит удаление незначительных деталей и разрешение задачи переобучения. Существуют субдискретизации не только с функцией максимизации, но еще и функцией среднего или L2 — нормирования.

5. Полносвязный слой(fully–connected layer)

   Данный слой представляет собой обычную нейронную сеть, где каждый нейрон связан со всем предыдущими. Предназначение такого слоя в том, чтобы вычислить уже по имеющимся данным вероятность принадлежности к классу. Значительно сократив объем данных по сравнению с исходными, наша сеть их передает в полносвязный нейронную сети, в которой может быть еще несколько слоев внутри.

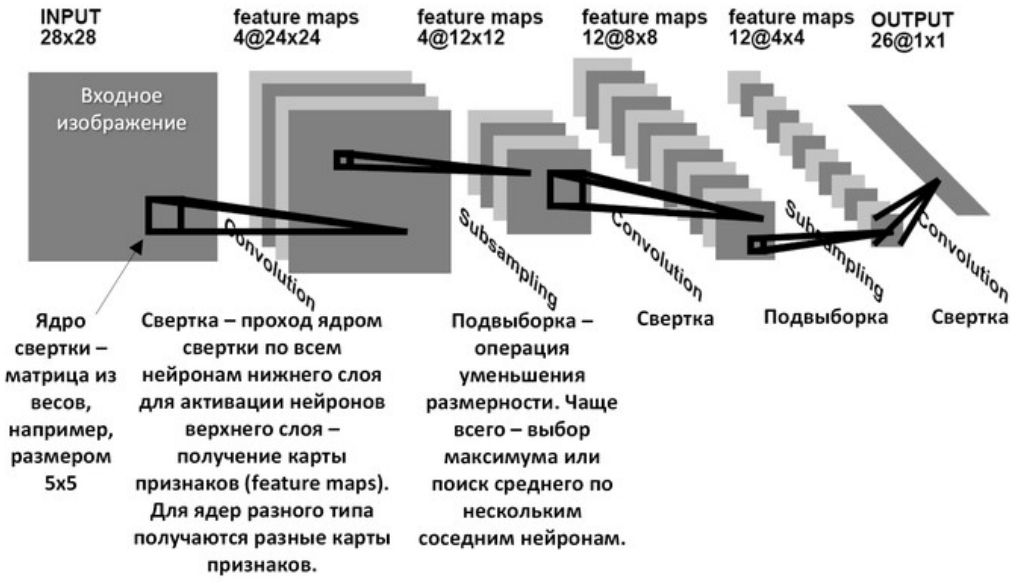


Рис. 9: Архитектура CNN

У сверточной нейронной сети [20] есть огромные преимущества в распознавании объектов:

- Высокая точность при распознавании
- Наилучший выбор при работе с изображениями
- Удобное распараллеливание, вследствие, возможность обучения и реализации на графическом процессоре
- Обучается такая сеть при использовании метода обратного распространения ошибки

Из недостатков можно смело выделить неопределенность структуры, многие параметры необходимо настраивать под задачу, то есть эмпирическим путем, методом проб и ошибок, а обучение сети на огромных датасетах может длиться неделями. Поэтому для многих задач берется уже обученная модель с матрицей весов, так как намного более перспективно дообучать сеть на своих данных. Особенно это актуально при небольшом количестве изображений. На данный момент существует множество архитектур, построенных на основе сверточных нейронных сетей. Примерами являются: AlexNet, VGGNet [21], ResNet [22], GoogLeNet.

# 3. Реализация алгоритма

## 3.1 База изображений

Первые проблемы, с которыми приходится столкнуться при обучении сети — большое количество шумов на изображениях, неоднородные данные или просто недостаточное количество картинок для полноценного обучения сверточной нейронной сети. Поэтому, чтобы избежать всех этих сложностей, мной была выбрана очень популярная база изображений с тегами, отображающими эмоции — fer2013dataset [23]. Данная выборка содержит в себе 28709 фотографий в тренировочном множестве и по 3589 в публичном и приватном множестве, соответственно. В конце 2013 года по этой базе проводилось ежегодное соревнование среди ученых и энтузиастов, занимающимися машинным обучение и анализом данных. Выбор пал именно на этот дата сет, так как входными данными являлась таблица со значениями пикселей в градации серого, вдобавок ко всему изображения были 48*48, что значительно уменьшало время обучения. Полноценная фотография 256*256 (пример с ImageNet) с RGB цветовой моделью представляла бы собой вектор в 86 больше, чем исходный вектор в моем алгоритме. Более того, в работе с этой базой данных не предусмотрена предобработка огромного количества изображений. Исследовав данную базу изображений можно заметить, что в ней недостаточно примеров фотографий с эмоцией «Отвращение», а именно 120 экземпляров, что явно недостаточно для полноценного обучения нашей нейронной сети. Поэтому мной было сделано стандартное изменение в исследованиях лицевых экспрессий посредством машинного обучения: замена в обучающей выборке «Отвращения» на «Злость», потому что с визуальной точки зрения данные эмоции близко эквивалентны. Тем самым получается обучающая выборка с примерно одинаковым распределением изображений с эмоциями. Выборка данных:

Таблица 1: Общее количество изображений в выборках

| Fer2013–Dataset | Тестовая выборка | Публичная | Приватная |
|---|---|---|---|
| Количество экземпляров | 28709 | 3589 | 3589 |

Таблица 2: Распределение эмоции на обучающей выборке

| | |
|---|---|
| Злость | 4431 |
| Радость | 7215 |
| Нейтрально | 4965 |
| Грусть | 4830 |
| Удивление | 3171 |
| Страх | 4097 |

## 3.2 Сверточная нейронная сеть

Как уже говорилось ранее, данная архитектура нейронной сети наиболее удобная и популяризированная для работы с изображениями. Поэтому очевидный выбор был сделан в пользу этой нейронной сети. Самая главная проблема сверточной сети [20] в том, чтобы подобрать необходимые параметры: количество слоев, необходимость MaxPooling и ZeroPadding. Более того существует необходимость в приведении сети к такому виду, чтобы сеть не переобучилась, и при этом выполнялось наилучшее соотношение времени точности. Первой архитектурой, которая была подобрана эмпирическим путем, оказалась сеть такого вида:

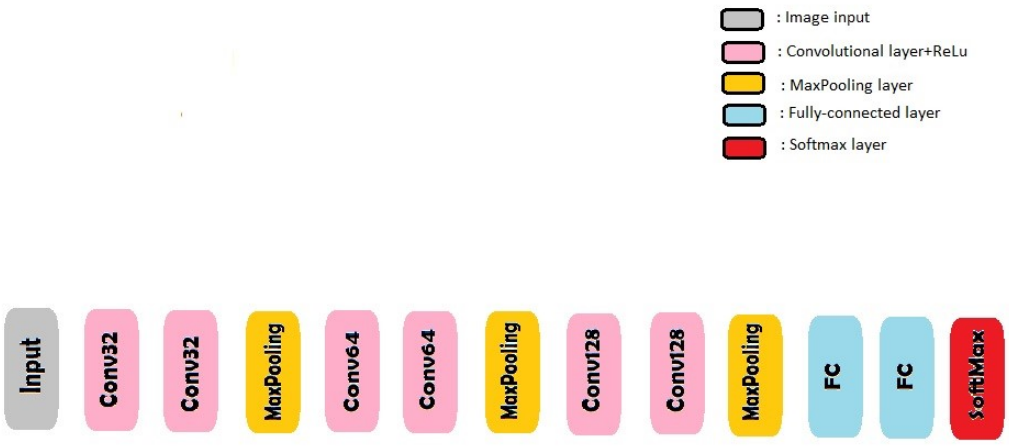


Рис. 10: Архитектура сверточной нейронной сети

Вторая архитектура, которую решено было использовать, была упрощенная модификация очень известной сети VGG–16 [21]. На основе такой архитектуры были выиграны ежегодные соревнования ImageNet ILSVRC–2014, где она показала наилучшую оценку в кластеризации и классификации. Для достижения наилучшего результата я решил сам обучать данную сеть на сервере с GPU, вместо того, чтобы брать уже натренированную сеть, и дообучить ее на имеющихся данных. Более того, можно было бы реализовать архитектуру более совершенных сетей, которые эффективно показали себя на практике, таких как: ResNet [22], GoogLeNet, но любой молодой исследователь в этой сфере упирается в проблему отсутствия необходимого железа для таких сетей.

## 3.3 Реализация

Для реализации системы распознавания эмоций и построения сверточной нейронной сети [18] [19] [20] был выбран Python. Данный язык программирования сейчас является одним из самых удобных в среде анализа данных и инструментов машинного обучения. Для обработки изображений было принято решение использовать популярную библиотеку OpenCV [13]. За построение модели нейронной сети отвечала библиотека с открытым

кодом Keras с предустановленной Theano. Это бесспорно самый удобная библиотека для работы с нейронными сетями и их конфигурацией.

Первой задачей, с которой необходимо было справиться было распределение имеющейся базы данных с изображениями, точнее их векторами значений пикселей, на тренировочную и тестовую выборку. В этой же программе существовала необходимость правильно распарсить таблицу и приравнять «Отвращения» и «Злость», чтобы на выходе получить два файла с выборками.

Обработав исходную таблицу и получив 2 выборки, необходимо начать реализацию самой архитектуры нейронной сети с библиотекой keras. Было принято решение о необходимости обучения итоговой сети на графическом процессоре (GPU). Потому что в сравнении с CPU, сверточная сеть обучается на GPU тренируется в 2500 раз быстрее. Такие данные были получены на CPU: Intel Core i5 (2.8 GHz) и GPU: Nvidia GeForce GTX 1080. Время на одну эпоху для первой архитектуры составляло 36 часов на процессоре и 47 секунд на графическом процессоре, соответственно. В связи с этим встала необходимость в аренде выделенного сервера с мощным GPU, так как работа с нейронными сетями очень энергоемкий процесс.

Выводом нашего обучения была матрица весов и архитектура моделей, сохраненные в файлах для дальнейшего использования или для дообучения на новом пуле данных.

Следующим важным шагом в построение нашей нейронной сети являлся проверка на тестовых данных. Интернет–ресурс Kaggle, устраивающий данное соревнование, после его окончания сделал доступным приватную тестовую выборку. Поэтому нейронная сеть была проверена на обоих множествах и были получены такие результаты:

Как видно из таблицы, точность алгоритма, полученная на тестовых данных очень близка к точности определения эмоций самим человеком на этой выборке. А это значение составляет порядка $65\% \pm 5\%$. То есть на такой сложной, комплексной и очень субъективной задаче машинное обучение достигает точности человеческого глаза при определении эмо-

Таблица 3: Точность нейронной сети на трех выборках

| | Обучающая выборка | Публичная тестовая выборка | Приватная тестовая выборка |
|---|---|---|---|
| Архитектура №1 | 64.2434% | 65.786% | 64.634% |
| Архитектура №2 | 65.967% | 65.432% | 66.230% |

ций. Далее следует работы алгоритма распознавания на тестовой выборке,представленный на Рис. 11, в котором можно наблюдать изображения лица человека и график вывода нормированного вектора вероятностей эмоций:

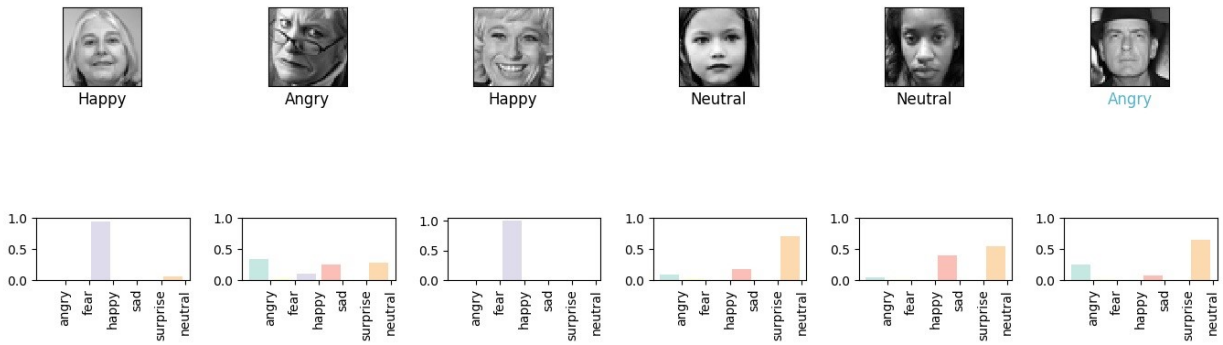


Рис. 11: Пример работы распознавателя

## 3.4 Использование алгоритма

Самой важной деталью программы и алгоритма является работа на реальных данных. Поэтому мной были взяты фотографии согласившихся на обработку данных студентов, которые должны были продемонстрировать на фото определенное проявление эмоций. Получился маленькая выборка, которая должна была проверить адекватность работы программы. Первое, что необходимо было сделать–это обработать полученную фотографию, ведь нейронная сеть принимает только фотографии размером 48*48

и в серой тональности. Поэтому с помощью инструментов OpenCV [13] в программе производятся необходимые преобразования, которые уже позволяют сделать прогноз при помощи уже обученной нейронной сети. Пример работы на реальных фотографиях:

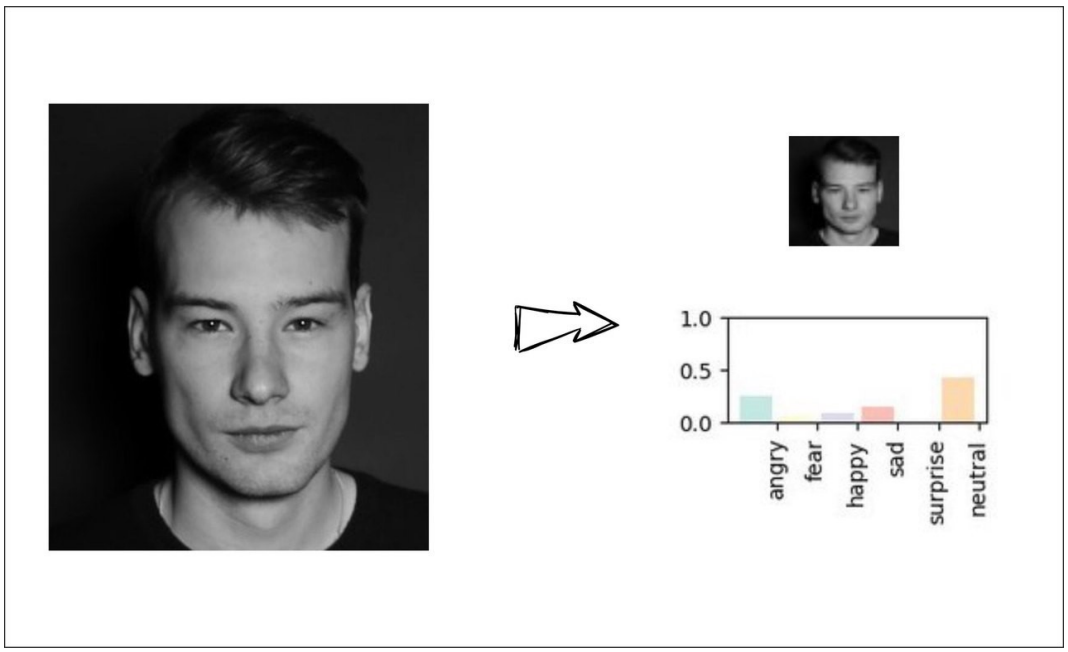


Рис. 12: Пример работы распознавателя с нейтральной эмоцией


Рис. 13: Пример работы распознавателя с счастливым лицом

# 4. Заключение

В проделанной работе была реализована программа для распознавания эмоций на человеческом лице по изображению. Был придуман алгоритм обработки входных данных, а также была построена архитектура нейронной сеть, которая успешно была протестирована на тестовых данных. Более того, полученная нейронная сети идеально удовлетворяла соотношению время обучения/точность алгоритма. Так как обучение нейронной сети–очень долгий и сложный процесс, поэтом данный фактор играл ключевую роль при выполнении работы. Самым трудным этапом в данной работе являлся эмпирический подбор параметров сети, который определял финальную точность. Получение весов при обучении сверточной нейронной сети–процесс на подобии «черного ящика», этим обуславливалась критическая проблема при построении такой архитектуры.

При выполнении практической и теоретической части работы были выполнены данные поставленные задачи:

1. Изучение такого инструмента машинного обучения как нейронные сети, которые в настоящее время превалируют в анализе данных и показывают наилучшие результаты со сложными задачами. На этом этапе так же были выделены и рассмотрены разные существующие методы, с помощью которых решалась задача детектирования лицевых экспрессий, был проведён сравнительный анализ и выявлены достоинства и недостатки каждого из методов.

2. Смоделировал архитектуру сверточной нейронной сети, которая показывает один из наилучших результатов точности детектирования эмоций. В дополнении к этому, обучение на такой сети оказалось оптимальным для доступных машинных ресурсов, и при этом обучалась в более 2000 раз быстрее, чем на процессоре.

3. Был успешно создан алгоритм по распознаванию эмоций на человеческом лице, а также программа была проверена в условиях реальной жизни. При несложных манипуляциях алгоритм такого типа можно

встраивать в системы видеонаблюдения и лайв–детектирования.

4. В работе, в первую очередь, была поставлена задача показать высокую точность при детектировании. Достигнутая точность оказалось на одном уровне с точностью определения эмоция человеком на данной выборке. Такие цифры были получены с помощью проведения эксперимента учеными, создавшими эмоциональную базу данных.

## 4.1 Перспективы

Завершая данную работу, было намечено огромное количество перспектив развития данной области и усовершенствования алгоритмов по распознаванию эмоциональных окрасок на человеческом лице при помощи средств машинного обучения. Прежде всего, нельзя забывать, что эмоции — очень субъективная окраска человеческого лица, очень часто даже человек не может определить лицевую экспрессию без комплекса других деталей, которые мы можем отмечать во время диалога с собеседником. Часто появляется необходимость в детектировании таких деталей, как тембр голоса, телодвижения и ситуационная составляющая. Ситуация бесспорно влияет на наше восприятие ситуации, являясь зачастую источником этих самых изменений в эмоциональной окраске. Поэтому в перспективе распознавания эмоциональной составляющей человеческого лица—создание комплекса алгоритмов, которые бы учитывали все тонкости. Голосовое детектирование вместе с трекингом изменений окружающей среды, распознавание основных психологических телодвижений индивидуума и все это в совокупности с обычным эмоциональным детектором порядком увеличили бы точность любых программных средств в этой проблеме машинного обучения. Более того, уже сейчас начинают появляться комплексные алгоритмы, которые используют параллельные нейронные сети, для вычисления ключевых точек и обычную сверхточную нейтронную сеть на основе ResNet уже обученной на ImageNet. Данные матрицы весов, полученные с помощью этих сетей, являются ещё одной тренировочный выборкой для объединяющей их обычный нейронной сети. Такие сложные и громоздкие

вычисления производятся в огромных дата центрах, при использовании современных вычислительных машин с параллельными графическими процессорами нового поколения.

Необходимость создания тенденции на интегрирование систем распознавания и систем видеонаблюдения уже становится понятной многим крупных компаниям, так как это влечёт за собой интерес государства и огромный финансовый успех на рынке систем безопасности. В дополнение ко всем, нельзя не упомянуть, что развитие таких систем неуклонно влечёт технологии и науку, в целом, к созданию искусственного интеллекта и любых его форм. Ведь умение понять, что чувствует собеседник или стопорный человек—одна из уникальных человеческих способностей, которая пока в полной мере недоступна компьютеру.

# Список литературы


[1] Darwin C., The Expression of the Emotions in Man and Animals, Oxford University Press, 1998. 472p

[2] Ekman P., Darwin and Facial expressions, Academic Press, 1973. 273p

[3] Ekman P., Facial expression and emotion, American Psychologist, 48:384-392, 1993

[4] Sebe N., Emotion recognition using a Cauchy Naive Bayes classifier // Pattern Recognition, 2002. Proceedings. 16th International Conference on, 10 December 2002. P. 34–46

[5] Nefian A., Hayes M., Hidden Markov model for face recognition // Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on, 06 August 2002. P. 48–56

[6] Li E., Wang B., GPU and CPU Cooperative Accelaration for Face Detection on Modern Processors // Multimedia and Expo (ICME), 2012 IEEE International Conference on, 13 September 2012. P. 67–89

[7] Esau N., Wetzel E., Real-Time Facial Expression Recognition Using a Fuzzy Emotion Model // Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International, 27 August 2007

[8] Kita S., Mita A., Emotion Identification method using RGB information of human face // Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems, 2015

[9] Agarwal M., Jain N., Face Recognition Using Principle Component Analysis, Eigenface and Neural Network // Signal Acquisition and Processing, 2010. ICSAP '10. International Conference on, 18 March 2010. P. 34–45

[10] Kar A., High Performance Human Face Recognition using Gabor Based Pseudo Hidden Markov Model // International Journal of Applied Evolutionary Computation (IJAEC), 2013. P. 11–22

[11] Srinivasan M., Vijayakumar S.,Pseudo 2D Hidden Markov Model Based Face Recognition System Using Singular Values Decomposition Coefficients // In The 2013 International Conference on Image Processing, Computer Vision,& Pattern Recognition(IPCV 2013), 2013. P. 252-258.

[12] Documentation of OpenCV. http://docs.opencv.org

[13] Gregori E. Introduction To Computer Vision Using OpenCV // Embedded Systems Conference in San Jose, 2012.

[14] Borylo P., Face Occurrence Verification Using Haar Cascades - Comparison of Two Approaches // Communications in Computer and Information Science, 2011. 149p

[15] Haar–like features. https://en.wikipedia.org/wiki/Haar-like_features

[16] Viola-Jones object detection framework.http://en.wikipedia.org/wiki/Viola-Jones/obhyject/detection/framework

[17] Freund Y., A short Introduction to Boosting // Journal of Japanese Society for Artificial Intelligence, 1999. P. 771-780

[18] Convolutional Neural Networks (CNNs / ConvNets). http://cs231n.github.io/convolutional-networks/

[19] Liu T., Fang S., Zhao Y., Wang P., Zhang J., Implementation of training convolutional neural networks // arXiv, 2015

[20] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. // Neural Information Processing Systems Foundation, Inc.: Ljubljana, Slovenia, 2012. P. 1097–1105.

[21] Simonyan K., Zisserman A., Very Deep Convolutional Networks for Large-Scale Image Recognition // International Conference on Learning Representations, 2014.

[22] Kaiming H., Xiangyu Z., Deep Residual Learning for Image Recognition // CVPR, 2016.

[23] Goodfellow I., Erhan D., Carrier P., Courville A., Challenges in representation learning: A report on three machine learning contests. Neural Networks, 64:59 – 63, 2015.