Комбинированный метод детектирования лиц для автоматической модерации пользовательских аватаров

Денис Тимошенко
Санкт-Петербургский государственный университет
Санкт-Петербург, Россия
e-mail: timoshenko.d.m@gmail.com

Валерий Гришкин
Санкт-Петербургский государственный университет
Санкт-Петербург, Россия
e-mail: valery-grishkin@yandex.ru

РЕЗЮМЕ
В настоящей работе описан способ решения задачи автоматической модерации графических представлений пользователей социальных сетей. Предлагается использовать комбинированный метод детектирования лиц на основе алгоритма Виолы-Джонса и сверточных нейронных сетей. Указанный подход позволяет достаточно точно производить подсчет лиц на графическом представлении пользователя, таким образом делая выводы о корректности используемого аватара.

Ключевые слова
Обработка изображений, машинное обучение, алгоритм Виолы-Джонса, вейвлеты Хаара, локальные бинарные шаблоны, сверточная нейронная сеть, деформируемые модели.

1. Введение
Одним из наиболее распространенных и важных пунктов пользовательского соглашения с социальными сетями является предоставление корректной информации о личности владельца аккаунта. Концепция социальной сети нарушается, когда становится невозможным установить внешний вид пользователя из-за отсутствия изображения лица на аватаре или наличия посторонних лиц. Ручная проверка всех графических представлений слишком трудоемка и экономически не выгодна для компании, обслуживающей социальную сеть. Поэтому, целесообразно применять автоматические методы обработки изображений, позволяющие с заданной точностью детектировать лица.

Предлагаемый в настоящей работе метод позволяет достаточно точно локализовать изображения лиц на фотографиях. Информация о количестве, размере и положении лиц может быть применена для автоматической фильтрации пользовательских аватаров.

2. Методы детектирования лиц
Под детектированием лиц понимается нахождение на пиксельной матрице областей, содержащих основные элементы лица. В рамках этой работы под границами лица мы будем подразумевать прямоугольную область изображения с вписанным овалом лица, исключая области шеи, ключиц и прическу.

Наилучшие показатели при распознавании лиц достигаются для идеального положения анфас, но на практике это довольно редкий случай. В этой работе для изображения лица считаются приемлемыми отклонения головы от положения анфас на +-45˚ в плоскости изображения и +-20˚ в остальных плоскостях. Профильные изображения лиц не рассматриваются.

2.1. Алгоритм Виолы-Джонса
Одним из самых популярных алгоритмов детектирования лиц является метод Виолы-Джонса [1]. Благодаря простоте реализации, скорости и высокому проценту обнаружения объектов он стал основой многих коммерческих детекторов.

Рис. 1. Пример вейвлетов Хаара.

Алгоритм базируется на идее последовательного построения композиции элементарных классификаторов. В качестве признаков изображения используются функции, подобные вейвлетам Хаара (рис. 1). Каждой функции соответствует "слабый" классификатор *H*:

$$H(x, f, p, q) = \begin{cases} 1, & \text{если } p * f(x) < p * \theta \\ 0, & \text{иначе} \end{cases}$$

В процессе обучения системы из всех возможных признаков выбираются наиболее подходящие для классификации. Выбор осуществляется с помощью алгоритма бустинга. Таким способом формируется набор наилучших "слабых" классификаторов, образующих один "сильный" классификатор.

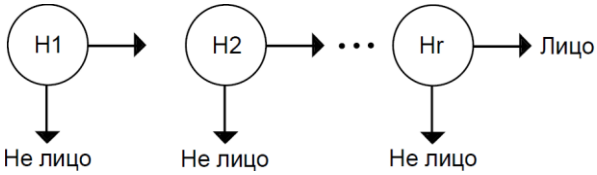

Рис. 2. Уровень каскада Виолы-Джонса.

Формируется вырожденное дерево решений, называемое каскадом. Каждый уровень каскада (рис. 2) состоит из "сильного" классификатора, обученного с помощью бустинга на ошибках предыдущего уровня.

Несмотря на высокую обобщающую способность алгоритмов бустинга и огромное количество классификаторов на практике главным недостатком реализаций алгоритма Виолы-Джонса является большой процент ложных срабатываний. Комбинированный подход позволяет воспользоваться высоким процентом обнаружения лиц, одновременно устраняя этот недостаток.

2.2. Локальные бинарные шаблоны

Помимо вейвлетов Хаара в качестве признаков для решающего дерева предлагается применять локальные бинарные шаблоны (ЛБШ) [2]. ЛБШ представляет собой описание окрестности пикселя изображения в двоичной форме. Оператор ЛБШ, который применяется к пикселю изображения, использует восемь пикселей окрестности, принимая центральный пиксель в качестве порога.

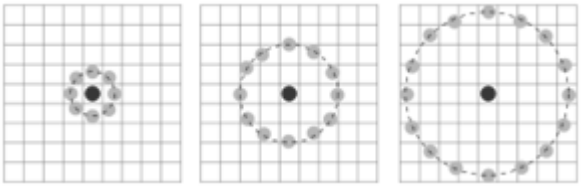

Рис. 3. Расчет локальных бинарных шаблонов.

Пикселям, которые имеют значения, большие чем центральный пиксель, сопоставляют единицы, а тем, которые меньше центрального, сопоставляют нулевые значения. Таким образом получается многоразрядный двоичный код, который описывает окрестность пикселя. Оператор может быть расширен и на большее число пикселей окрестности (12, 16 и более) с увеличением радиуса окрестности, как показано на рис 3.

2.3. Сверточная нейронная сеть

При решении задач обработки зрительной информации и инвариантного распознавания изображений интенсивно развивается подход, основанный на разработке вычислительных алгоритмов, имитирующих принципы работы реальных зрительных систем (бионический подход), который рассматривается как наиболее перспективный.

В середине прошлого столетия ученые Torsten Nils Wiesel и David Hunter Hubel исследовали зрительную кору головного мозга кошки и обнаружили, что существуют так называемые простые клетки, которые особо сильно реагируют на прямые линии под разными углами и сложные клетки, которые реагируют на движение линий в одном направлении [3].

Позже, исследователь Ян ЛеКун (Yann LeCun) предложил использовать так называемые сверточные нейронные сети (СНС) для решения задачи распознавания образов [4]. СНС, внедренные и успешно использованные ЛеКуном, это мощные бионические иерархичные многослойные нейронные сети, которые объединяют три архитектурные идеи, чтобы обеспечить некоторую степень сдвига, масштаба и инвариантности представления: локальные рецептивные поля, общие веса и пространственная субдискретизация. Различные архитектуры СНС успешно используются во многих сложных приложениях, например таких как распознавание рукописного ввода или классификация изображений.

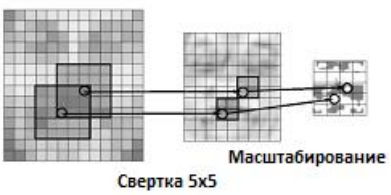

Рис. 4. Операции СНС.

К входному изображению, которое располагается на входном слое-ретине, последовательно применяются операции свертки и субдискретизации (рис. 4). После ряда чередующихся сверточных и масштабирующих слоев, как правило, располагается полносвязный перцептрон.

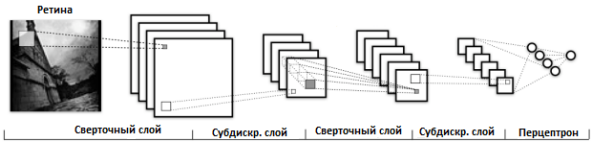

Рис. 5. Структура LeNet5

В настоящей работе применялась СНС, повторяющая структуру сети LeNet5 [4]. Параметры выбирались как у модифицированной сети LeNet5, описанной Гарсией и Делакисом (Garcia and Delakis) в работе [5]:

- Размер ретины 32x32 пикселя.
- Количество нейронов первого сверточного слоя равнялось 4, размер ядра свертки – 5.
- Количество нейронов второго сверточного слоя равнялось 14, размер ядра свертки – 3.
- Масштаб карт в субдискретизирующих слоях уменьшался в два раза.

2.3. Модель деформируемых частей

В рамках бионического подхода особое внимание уделяется разработке алгоритмов и методов определения наиболее информативных областей изображений для детальной обработки, как аналогов биологических механизмов выбора перцептуально важных фрагментов при осмотре изображений. Одним из самых первых и основных методов является анализ геометрических характеристик лица. Данный метод предъявляет более строгие требования к условиям съёмки, нуждается в надёжном механизме нахождения ключевых точек для общего случая. Кроме того, требуется применение более совершенных методов классификации или построения модели изменений.

Модель деформируемых частей (МДЧ) – один из методов, позволяющих с высокой точностью определять расположение ключевых областей на изображении лица. МДЧ представляет лицо как набор моделей внешнего вида различных его частей, местоположение которых определяется заданной конфигурацией (графом). В дополнение к указанным частичным моделям, одна используется для представления внешнего вида всего лица, независимо от взаимного положения его частей.

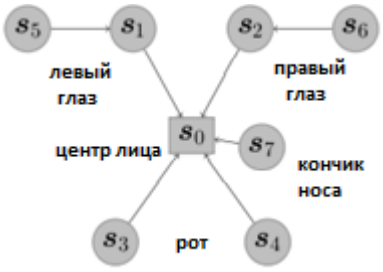

Рис. 6. Конфигурация МДЧ для лица.

Каждая из этих моделей в качестве признаков использует гистограммы ориентированных направлений (HOG).

Оценочная функция определяется как сумма меры внешеней схожести частей с их обученными моделями и функции стоимости деформации графа, связывающего отдельные модели:

$$f(I,s) = \sum_{i=0}^{7} q_i(I,s_i) + \sum_{i=0}^{4} g_i(s_0,s_i) + g_5(s_1,s_5) + g_6(s_2,s_6) + g_7(s_0,s_7)$$

здесь s – модель, соответствующая определенной части лица(см. рис. 6), I – исходное изображение, q и g – мера схожести и функция стоимости соответственно [6].

Для обучения параметров функций обычно применяют какую-либо из модификаций машины опорных векторов.

Рис. 7. Пример выделения ключевых точек.

На рис. 7 приведен пример работы алгоритма МДЧ, реализованного в библиотеке с открытыми исходными кодами flandmark [7]. Белым цветом отмечены точки, означающие соответствующие модели частей (границы глазных щелей, уголки рта, нос и центр лица).

2.3. Комбинированный метод

Основной идеей комбинированного подхода является применение фильтра к результатам детектирования серии решающих деревьев на основе ЛБШ и вейвлетов Хаара. Серия детекторов включает в себя деревья решений, обученные на различных наборах данных, различающихся углом поворота лица в плоскости изображения. Как показывает практика, методы Виолы-Джонса и ЛБШ сохраняют обобщающую способность в диапазоне 20-25˚ поворота головы.

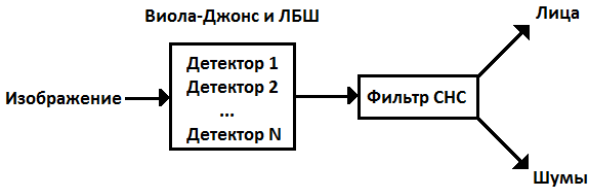

Рис. 8. Схема комбинированного метода.

В качестве фильтра предлагается использовать СНС, обученную на два класса: “лицо” и “шум”. Стоит отметить, что данный подход отличается от способа, описанного в работе [5] именно тем, что СНС не приходится сканировать изображение целиком, поскольку эту задачу решают детекторы, описанные в разделах 2.1 и 2.2.

3. Обучение алгоритмов

Обучающая выборка была сформирована из следующих баз: BioID, FERET, Caltech, ORL database, Georgia Tech Face Database, Sheffield Face Database, CVSRP. Отдельно была собрана база фотографий из открытых альбомов в социальных сетях. Общее количество изображений превышало 10 тыс., количество лиц на изображениях – 12 тыс.

Все базы содержали в основном лица, удовлетворяющие приемлемым отклонениям. Более того, исключались изображения, на которых лицо было закрыто каким-либо предметом более чем на треть.

3.1. Подготовка данных

С целью усилить обобщающую способность алгоритмов на классе лиц к исходным фотографиям из базы обучения выборочно применялась серия трансформаций: повороты на +-20˚ и +-45˚ в плоскости изображения, размытие ядром Гаусса со стороной 3 и 5, двукратное снижение и увеличение яркости. Размытие и изменение яркости также применялись и к повернутым изображениям. Это позволило расширить обучающую выборку почти до 30 тыс. изображений.

В эксперименте, описанном в работе [5], производилось выравнивание изображений лиц на ретине. В качестве

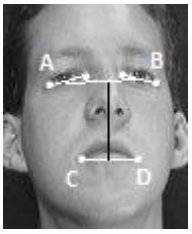

Рис. 8. Линии выравнивания изображений лиц.

ориентира использовалась линия AB, соединяющая центры глаз, и расстояние от этой линии до линии губ CD (рис. 8). Авторы производили ручную разметку положений глаз и губ почти на 4 тыс. изображений.

С целью экономии времени на подготовку базы, было применено автоматическое сегментирование на основе метода деформируемых моделей. Для этого была разработана утилита, включающая стандартный детектор лиц из пакета OpenCV [8] и сегментатор ключевых точек из библиотеки Flandmark.

Все выделенные фрагменты, на которых успешно были найдены ключевые точки, перед занесением в класс "лица" обрабатывались по следующему правилу:

1. По всей базе обучения рассчитывались следующие величины: d_M – среднее расстояние от AB до верхней границы лица; d_N – среднее расстояние от линии AB до CD.
2. На каждом изображении выделенная детектором граница лица смещалась таким образом, чтобы линия глаз оказалась ровно посередине относительно вертикальных сторон и на расстоянии в d_M пикселей от верхней границы.
3. Размер изображения изменялся на коэффициент, равный отношению расстояния от AB до CD к величине d_N.
4. К изображению применялась упомянутая выше серия трансформаций: повороты, размытие, изменение яркости.
5. Затем полученные изображения обрезались по контуру границы лица, и выделенная область масштабировалась к размеру 32x32 пикселя.

Прочие фрагменты изображений, не содержащие лиц или содержащие какие-либо отдельные части были отнесены к классу "шумы".

3.2. Обучение каскадов

Деревья решений на вейвлетах Хаара и ЛБШ обучались с помощью утилиты *cascadetrain* программного пакета OpenCV. Было сформировано пять групп изображений из базы обучения, на которых раздельно тренировались пять классификаторов. Параметры обучения в основном не выбирались специальным образом. Ставилась цель обучить классификаторы с максимально возможным параметром HR (Hit Rate, коэффициент попаданий). Использовались различные признаки для исходных изображений базы, и изображений, повернутых на определенный угол на этапе подготовки данных. Величина FAR (False Acceptance Rate, уровень ложных принятий) регулировалась, исходя из поведения процесса обучения: слишком малые величины приводили к деградации HR.

Таблица 1

Параметры обучения каскадов

№	Признаки	Угол	Бустинг	HR	FAR
1	Хаара	0	LogitBoost	0,998	0,5
2	ЛБШ	+20˚	GentleBoost	0,995	0,2
3	ЛБШ	+45˚	GentleBoost	0,995	0,2
4	ЛБШ	-20˚	GentleBoost	0,995	0,2
5	ЛБШ	-45˚	GentleBoost	0,995	0,2

3.3. Обучение сверточной сети

Обучение СНС осуществляется с помощью алгоритма обратного распространения ошибки. Обучающие данные обрабатывались пакетами по 10 изображений. На каждой итерации данные перемешивались случайным образом и перераспределялись.

Для обучения СНС были сформированы вручную два набора данных. Первый составляли изображения лиц, полученные путем обработки обученными каскадами из 3.2 расширенной обучающей базы. Второй – ошибки второго рода серии каскадов, т. е. различные шумы.

Использовался следующий пошаговый алгоритм обучения с переменным объемом обучающей базы на каждой серии итераций:

1. Инициализируется начальная выборка для классов лиц и шумов, *ecount=30, inum=20.*
2. Начальная выборка устанавливается в качестве обучающей.
3. Осуществляется тренировка сети на обучающей выборке в течение *ecount* эпох.
4. Полученная сеть тестируется на базе обучения; изображения, на которых сеть ошибается, составляют новую выборку.
5. Новая выборка устанавливается в качестве обучающей, переменная *inum* уменьшается на единицу. Если *inum* больше нуля, то возвращаемся к шагу 3. Иначе – окончание процесса обучения.
6. Согласно схеме алгоритма совокупное число эпох обучения СНС равно 600. Кроме того, на последующих итерациях, когда эмпирический риск сети уменьшается, скорость обучения как правило возрастает.

3. Результаты

Тестирование системы проводилось на базе Labeled Faces In Wild (LFW) [9] и наборе фотографий, собранных из социальных сетей авторами статьи (Social1). В базу Social1 были включены большей частью сложные для детекторов условия: значительные повороты голов, расфокусировка и групповые фото низкого разрешения. Для обеих баз была подготовлена экспертная эталонная разметка.

Таблица 2

Результаты тестирования методов

База	Кассификаторы	Recall, %	FAR, %
LFW	Хаара+ЛБШ	72	33,4
Social1	Хаара+ЛБШ	67,02	43,6
LFW	Комбинированный	70,1	9,4
Social1	Комбинированный	55,3	4,92

В табл. 2 приведены результаты тестирования двух методов: детекторов Виолы-Джонса на признаках Хаара вместе с решающими деревьями на ЛБШ и комбинированного. Проценты обнаруженных лиц считались по правилу, установленному в состязании LFW: если площадь пересечения эталонной разметки и тестовой превышала половину площади объединения двух разметок, то лицо считалось верно найденным.

4. Заключение

Предложен новый комбинированный метод обнаружения лиц на статических изображениях с использованием сверточных нейронных сетей, алгоритма Виолы-Джонса и ЛБШ. Метод детектирует большую часть лиц при достаточно малом проценте ложных срабатываний. Данные свойства метода позволяют применять его в задачах обработки фотографий социальных сетей. Путем выставления порогов фильтра можно варьировать лояльность системы модерации к пользователю и количество пропущенных некорректных фотографий.

REFERENCES

[1] P. Viola, M. J. Jones "Robust Real-Time Face Detection", *International Journal of Computer Vision*, pp. 137-154, 2004.

[2] Xiaoyu Wang, Tony X. Han, Shuicheng Yan, "An HOG-LBP Human Detector with Partial Occlusion Handling", *International Conference on Computer Vision*, pp. 32-39, 2009.

[3] D. H. Hubel, Wiesel T. N. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex", *Journal of Physiology*, pp. 106-154, 1962.

[4] Y. LeCun, L. Bottou, G. Orr, K. Mller. "Effcient BackProp. In Neural Networks: Tricks of the trade", *Springer Lecture Notes in Computer Science*, pp.5-50, 1998.

[5] C. Garcia, M. Delakis "Convolutional face finder: a neural architecture for fast and robust face detection", *Pattern Analysis and Machine Intelligence*, pp. 1408-1423, 2004.

[6] M. Uricar, V. Franc and V. Hlavac, "Detector of Facial Landmarks Learned by the Structured Output SVM", *Proceedings of the 7th International Conference on Computer Vision Theory and Applications*, pp. 547-556, 2012.

[7] Flandmark, http://cmp.felk.cvut.cz/ ~uricamic/flandmark/

[8] OpenCV, http://opencv.willowgarage.com/wiki/

[9] LFW database, http://vis-www.cs.umass.edu/lfw/