

Институт прикладной математики имени М. В. Келдыша
Российской академии наук

На правах рукописи

Конушин Вадим Сергеевич

**Исследование и разработка алгоритмов
анализа аудитории по данным видеонаблюдения**

Специальность 05.13.11 – математическое и программное обеспечение
вычислительных машин, комплексов и компьютерных сетей

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва – 2014

Работа выполнена в Институте прикладной математики им. М.В.Келдыша РАН

Научный руководитель: кандидат физико-математических наук, доцент
Баяковский Юрий Матвеевич

Официальные оппоненты: _____.

_____.

Ведущая организация: _____.

Защита состоится «_____» 2014 г. в _____ часов на заседании диссертационного совета
по адресу: _____.

С диссертацией можно ознакомиться в библиотеке _____.

Автореферат разослан «_____» _____ 2014 г.

Учёный секретарь диссертационного совета

_____.

Общая характеристика работы

Объект исследования и актуальность работы

Анализ данных видеонаблюдения или видеоаналитика – одна из наиболее востребованных, а потому и одна из наиболее активно исследуемых областей компьютерного зрения. Установка камер видеонаблюдения в последние годы стала практически повсеместной. Камеры установлены в торговых и бизнес центрах, ресторанах и банках, аэропортах и вокзалах, большинстве школ и во многих подъездах жилых домов. Также камерами оснащены все банкоматы, целый ряд моделей телевизоров, множество персональных компьютеров и ноутбуков. С удешевлением жестких дисков постепенно исчезает проблема хранения больших архивов видео. Однако обработка и анализ видео по большей части выполняется по-прежнему вручную: либо в режиме реального времени за несколькими каналами видео наблюдает оператор, либо уже постфактум осуществляется последовательный просмотр всех записей в видеоархиве для поиска нужных событий, например, кражи на объекте наблюдения. Поэтому существует высокий спрос на автоматизацию обработки и анализа данных видеонаблюдения.

К сожалению, существующие алгоритмы компьютерного зрения позволяют решать только отдельные задачи видеоаналитики для узких сценариев использования. В качестве примера можно привести задачу подсчета числа людей в магазинах по видео. С обычного ракурса видеонаблюдения, когда камера смотрит на входящих людей под углом, не удастся точно посчитать число людей из-за частых перекрытий. Поэтому в последние годы получают активное распространение системы¹², использующие потолочные камеры, которые направлены практически вертикально вниз. В таком ракурсе, при наличии контрастной относительно пола одежды люди становятся хорошо разделимыми, и удастся достигнуть высокой точности подсчёта их количества. Другими примерами решенных задач являются распознавание номеров машин, или сценарий «стерильной зоны» для обнаружения движения в запрещенной зоне. Такие же общие задачи, как сопровождение движущегося человека в толпе или идентификация человека по изображению лица с произвольного ракурса, в настоящий момент не решены.

Одной из актуальных задач является анализ аудитории по данным видеонаблюдения. Под таким анализом понимается подсчет числа людей, оказавшихся в поле наблюдения камеры и определение пола, возраста и других характеристик наблюдаемых людей.

¹<http://подсчет-посетителей.пф>

²<http://synesis.ru/solutions/cassiopeia/podschet-posetitelej-v-torgovyix-setyax-i-bankax>

Такой анализ важен, например, для сбора объективной информации об аудитории магазинов или ресторанов, отслеживания динамики аудитории в течение времени. Как и в ранее приведённых примерах, решение этой задачи в общем случае, при произвольном ракурсе, пока является слишком сложным. Поэтому естественно наложение ограничений на возможные ракурсы и сценарии съемки.



Рисунок 1. Иллюстрация рассматриваемого сценария видеонаблюдения.

В данной работе рассматривается сценарий видеонаблюдения, при котором камера закреплена на стене или стенде рядом с монитором внутри помещения, на высоте 1-2 метра от пола или земли, и наблюдает за людьми, подходящими к монитору (рис. 1). При таком положении камеры лица людей наблюдаются с близкого к фронтальному ракурсу, и их изображения получаются более чёткими, чем при закреплении камеры наблюдения под потолком, что облегчает задачу анализа изображений лиц. Такой сценарий подходит для анализа зрителей информационных или рекламных мониторов (Digital Signage), пользователей банкоматов, и т.д.

Разрешение видеопотока должно быть таким, чтобы обеспечивать размер лиц людей не ниже 60x60 пикселей, необходимых для определения характеристик человека по изображению лица. Резкие движения объектов в кадре затрудняют их сопровождение, поэтому частота кадров видеопоследовательности должна быть не ниже 15 кадров в секунду.

Для решения задачи анализа аудитории предлагается использовать стандартную для видеоаналитики схему решения задачи, состоящую из следующих этапов: выделения маски переднего плана (вычитание фона), выделения лиц людей на кадрах, сопровождения людей в видео, и классификации людей по изображению лица. В качестве алгоритма

выделения лиц было решено использовать известный алгоритм Viola-Jones³, для остальных же трех задач в рамках диссертации разработаны новые алгоритмы.

Цель диссертационной работы

Целью работы является разработка комплекса алгоритмов для анализа аудитории по видеопотоку с камеры видеонаблюдения, включающего алгоритмы решения таких задач, как вычитание фона, сопровождение людей в видеопотоке и классификация людей по изображению лица. Разработанные алгоритмы должны превосходить существующие по точности.

Научная новизна работы

В диссертации предложен новый алгоритм вычитания фона на основе поблочных классификаторов. Кадры видео разбиваются на набор непересекающихся блоков, для каждого из которых обучается отдельный классификатор, запоминающий текстуру фона в данном блоке. Благодаря такому обучению удастся устойчиво отличать перекрытие блока от изменения освещения. Также предлагается модификация алгоритмов вычитания фона для рассматриваемого ракурса наблюдения, на основе модели маски переднего плана, характерной для изображений людей при таких ракурсах. Предложенная модификация может использоваться не только для предложенного поблочного алгоритма, но и для других алгоритмов вычитания фона, и обеспечивает повышение точности без заметного прироста вычислительной сложности.

Разработан новый алгоритм сопровождения людей в видеопотоке, основанный на методе фильтра частиц. Предлагается явное моделирование состояния видимости объекта, благодаря которому алгоритм способен обрабатывать сложные случаи перекрытия между людьми. Также предлагается модификация, позволяющая корректно обрабатывать случаи одновременного входа и выхода двух человек.

Разработана новая схема алгоритмов классификации людей по лицу. Вначале оценивается сходство между тестовым лицом и лицами из известной базы, для чего обучаются специальные классификаторы сравнительных признаков внешности. Сама классификация выполняется на основе посчитанного сходства с этими людьми. Предложенный под-

³ P. Viola, M.J. Jones, Robust Real-Time Face Detection, *Int'l J. Computer Vision*, Vol 57(2), pp. 137-154,

ход был протестирован на задаче классификации пола человека и показал высокую точность классификации на эталонной коллекции LFW⁴.

Обоснованность и достоверность результатов

Достоверность результатов подтверждается проведённой экспериментальной оценкой всех предложенных алгоритмов. Для оценки алгоритмов вычитания фона и сопровождения людей в видео использовалась собственная выборка видеопоследовательностей, а также публично доступный набор ChangHai⁵. Для сравнения использовались авторские и собственные реализации алгоритмов вычитания фона, а также открытые реализации из библиотеки OpenCV. Для оценки алгоритма классификации людей использовалась эталонная коллекция LFW, на которой предложенный алгоритм показал одно из самых высоких на момент работы значений точности в 92,96%.

Практическая значимость и реализация

В рамках работы выполнена программная реализация разработанных алгоритмов в программной среде Matlab, удовлетворяющая всем требованиям и ограничениям, сформулированным при постановке задач. Разработанные алгоритмы могут быть использованы в системах видеонаблюдения, анализа аудитории, системах безопасности, индексирования и поиска по видеоархивам.

Апробация работы

Результаты работы докладывались и обсуждались на:

- 20-ой международной конференции по компьютерной графике и машинному зрению «Graphicon'2010», Россия, Санкт-Петербург, 2010;
- Международной конференции по компьютерному зрению DAGM (Deutsche Arbeitsgemeinschaft für Mustererkennung), Германия, Франкфурт, 2011;
- 21-ой международной конференции по компьютерной графике и машинному зрению «Graphicon'2011», Россия, Москва, 2011;
- 22-ой международной конференции по компьютерной графике и машинному зрению «Graphicon'2012», Россия, Москва, 2012;
- 3-ей конференции «Intelligent Video 2.0», Россия, Москва, 2012;
- Научной конференции «Ломоносовские чтения», Россия, Москва, 2013;

⁴ Huang G., Ramesh M., Berg T., Learned-Miller E. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments // UMass Amherst Technical Report 07-49, 2007.

⁵<http://www.cs.utexas.edu/~changhai/icra10-datasets/datasets.html>

- Семинаре по компьютерной графике и машинному зрению Ю.М. Баяковского (ф-т ВМК МГУ).

Публикации

По теме диссертации автором опубликовано 6 научных работ, из них 3 статьи в рецензируемых журналах, включенных в перечень ВАК РФ [3-4],[6] и 3 статьи в сборниках трудов международных конференций [1-2],[5].

Структура и объем работы

Диссертация состоит из введения, трех глав, заключения и списка литературы. Содержание работы изложено на XX страницах. Список литературы включает XX наименования. В работе содержится XX рисунка и XX таблицы.

Содержание работы

Во введении описывается общая задача анализа аудитории по данным видеонаблюдения. Т.к. на сегодняшний день данная задача далека от своего решения, формулируется более узкий целевой сценарий видеонаблюдения. В этом сценарии камера закреплена внутри помещения на стене или стенде рядом с монитором, на высоте 1-2 метра от пола или земли, и наблюдает за людьми, подходящими к монитору. Требуется определить число людей подошедших людей, а также классифицировать их по полу или другим характеристикам.

Описана схема решения данной задачи, состоящая из 4 последовательных этапов: выделения объектов переднего плана, выделения лиц людей, сопровождения людей в видео и классификации людей по изображению лица. Для выделения лиц людей используется известный алгоритм Viola-Jones. Для остальных шагов разработаны новые алгоритмы.

В первой главе рассматривается задача выделения переднего плана в видео. На входе алгоритм получает последовательные кадры видео $I_1, \dots, I_k$, снятые со стационарной камеры видеонаблюдения. Первые n кадров не содержат переднего плана и служат для настройки алгоритма. Требуется провести бинарную попиксельную разметку каждого кадра I_t на передний план F_t и фон B_t , т.е. построить бинарную маску M_t , в которой пикселям переднего плана сопоставлена метка 1, а пикселям фона – метка 0. Т.о. $I_t = M_t \cdot F_t + (1 - M_t) \cdot B_t$, где под “ $\cdot$ ” понимается поэлементное (попиксельное) умножение. Под фоном понимается совокупность объектов, которые присутствовали на первых n кадрах видео, а под передним планом - любые объекты, такие как люди, появившиеся хотя бы

на одном кадре после n -ого. Полученная маска переднего плана используется для сопровождения объектов в видеопотоке.

Ключевым элементом любого алгоритма вычитания фона является модель фона. Обычно модель фона инициализируется для каждого пикселя изображения по нескольким первым кадрам. Затем каждый пиксель последующего кадра сравнивается с моделью фона в данном пикселе. Если пиксель удовлетворяет модели фона, то он считается фоном, иначе – передним планом. После этого модель фона обновляется для пикселей, классифицированных как фон. В самом простом варианте в качестве модели фона можно использовать яркость каждого пикселя, а для сравнения использовать модуль разницы, отсюда и произошло название – «вычитание фона».

Критериями качества алгоритмов вычитания фона являются качество сегментации, скорость работы и требуемый объем памяти. При оценке качества сегментации переднего плана обычно отдельно считают попиксельные точность и отклик (полноту) маски переднего плана.

Для рассматриваемого сценария видеонаблюдения наиболее актуальной проблемой методов вычитания фона является изменение освещения, которое может происходить из-за плавного изменения освещения из окон, включения/выключения дополнительного света, изменения конфигурации источников света (например, открытие двери в комнату с ярким освещением), автоподстройки камеры. Одним из способов частичного решения данной проблемы является использование пространственной информации, например, построение моделей изменения освещения всей сцены или использование глобальных методов на основе Марковских случайных полей. Однако такие подходы имеют целый ряд недостатков, и существенно уступают попиксельным и поблочным методам по производительности. Другим способом является использование признаков локальной текстуры изображения, за счёт чего можно повысить устойчивость к изменению освещения.

В данной работе предлагается новый алгоритм вычитания фона, который сочетает устойчивость к изменениям освещения с высокой скоростью работы. Идея алгоритма заключается в разбиении кадра на набор непересекающихся блоков, для каждого из которых обучается отдельный бинарный классификатор {объект, фон}. Для обеспечения низкой вычислительной сложности предлагается использовать в качестве классификатора взвешенную линейную комбинацию простых классификаторов, которая строится с помощью метода «бустинга» - усиления слабых классификаторов. Для настройки алгоритма машинного обучения необходима эталонная выборка положительных и отрицательных примеров. Обычно эталонная выборка создается вручную оператором. Поскольку необходимость подготовки оператором эталонной выборки для каждого отдельного видеопотока не

позволит применять алгоритм на практике, предлагается генерировать такую выборку автоматически по одному или нескольким кадрам без объектов переднего плана, имитируя возможные изменения в освещении и настройках видеокамеры для положительных примеров, и возможные перекрытия фона для отрицательных примеров.

Таким образом, в предложенном алгоритме можно выделить следующие шаги:

1. Построение обучающей выборки
2. Обучение классификаторов на выборке
3. Использование обученных классификаторов для получения маски объектов переднего плана в видеопоследовательности.

Для обеспечения низкой вычислительной сложности классификатора блока {объект, фон} в качестве признаков предлагается использовать сравнения двух пикселей по цвету. Благодаря тому, что признак использует не абсолютное значение цвета, а его значение относительно цвета другого пикселя, то признак оказывается более устойчивым к изменениям освещения сцены. Поскольку всевозможных комбинаций двух пикселей для блока $N \times N$ пикселей $(N \times N)^2$, то для выбора наиболее информативных признаков предлагается использовать метод бустинга. Дополнительно, чтобы не выбирать пары соседних пикселей, а также для уменьшения исходного вектор-признака блока, блок разбивается на набор непересекающихся ячеек размера $M \times M$ (во всех экспериментах $M = 3$), после чего для дальнейшего рассмотрения из каждой ячейки остается только 1 пиксель (если N не делится на M , то ячейки в последнем ряду и столбце получают другого размера).

В качестве простых классификаторов используется решающее дерево глубины 2, в каждом узле которого сравнивается разница соответствующих цветовых каналов двух пикселей с порогом $C_{x,i} - C_{y,i} > t$, где $C_{x,i}$ - i -ый цветовой канал в пикселе x .

Итоговый классификатор блока является взвешенной комбинацией простых классификаторов: $a(x) = \text{sign}(\sum_{i=1}^K w_i b_i(x) - T)$, где w_i - вес i -ого признака, $b_i(x)$ - бинарный выход $\{0, 1\}$ решающего дерева, K - число простых классификаторов, T - порог. Меняя порог T , можно изменять соотношение отклика (полноты) и точности сегментации. Число простых классификаторов K по результатам экспериментов было выбрано равным 20.

Обучающая выборка для каждого классификатора должна состоять из набора положительных и отрицательных примеров для фоновое изображение. Под положительными примерами в данном случае понимаются изображения фона без перекрытия. Под отри-

цательными примерами понимаются изображения объектов переднего плана, в которых доля видимого фона составляет менее 40%.

Для создания положительной выборки к имеющимся фоновым изображениям добавляются примеры, полученные путем изменения их освещения, контрастности и небольшим размытием. Если среди исходных фоновых изображений есть примеры с различным освещением, то с помощью метода главных компонент можно получить усредненное фоновое изображение, и, добавляя к нему главные компоненты с разными коэффициентами, можно более правильно промоделировать разные варианты освещения сцены.

Для построения отрицательной выборки было решено моделировать «наложение» объектов переднего плана на фоновое изображение. Т.к. целевыми объектами переднего плана являются люди, то было решено ограничиться наложением изображений людей. Для этого была размечена маска людей на 20 изображениях, снятых с подобных ракурсов, с помощью интерактивного приложения сегментации изображений GrowCut⁶. Для упрощения сегментации люди снимались в основном на однородном контрастном фоне.

Определение переднего плана на каждом кадре видеопотока представляет собой применение классификатора {объект, фон} к каждому блоку. После этого применяется пространственная фильтрация с помощью медианного фильтра по окрестности 3x3. Стоит отметить, что т.к. фильтрация осуществляется на уровне блоков, а не пикселей, то время на её работу очень мало.

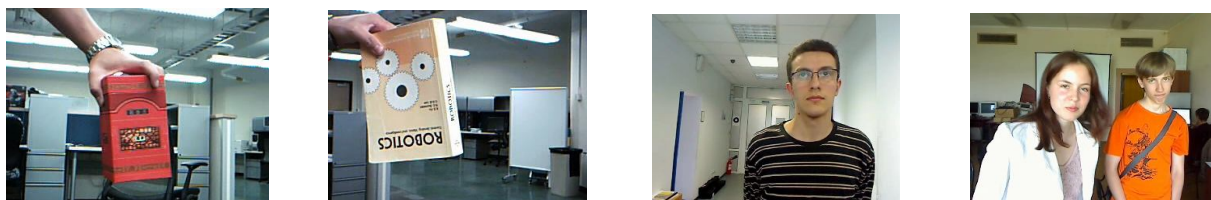


Рисунок 2. Примеры кадров из тестовых видеопоследовательностей.

Экспериментальная оценка разработанного алгоритма была проведена на 2 наборах видеопоследовательностей. Первый набор был подготовлен в Университете Техаса и содержит 3 видеопоследовательности, снятых внутри помещения. Набор был выбран для тестирования потому, что фон сцены содержит ряд объектов с характерной структурой, и во время видеопоследовательности несколько раз меняется освещение сцены, а также настройки камеры. Объектом переднего плана является коробка, которую подносят близко к камере. Имеется эталонная разметка каждого 5-го кадра. Вторым набором является собственной выборкой из 35 видеороликов, снятых с учетом рассматриваемого сценария ви-

⁶ <http://growcut.com/>

деонаблюдения. Эталонная разметка создана вручную для 150 случайных кадров. Примеры кадров из обоих наборов показаны на рис. 2.

Для сравнения использовались: один из наиболее распространенных алгоритмов вычитания фона на основе смеси Гауссиан (реализация, встроенная в Matlab); алгоритм Vibe, один из лучших на момент проведения эксперимента (реализация из OpenCV, а также собственная реализация); алгоритм на основе текстурных признаков (авторская реализация)⁷; алгоритм на основе Марковских случайных полей (собственная реализация)⁸.

Предложенный алгоритм показал в среднем наибольшую точность среди всех сравниваемых алгоритмов (см. рис. 3). Преимущество обеспечивалось в основном за счёт кадров после изменения освещения или настроек камеры, а также за счёт кадров, когда человек долго стоит почти без движения и уже встраивается в фон другими алгоритмами.

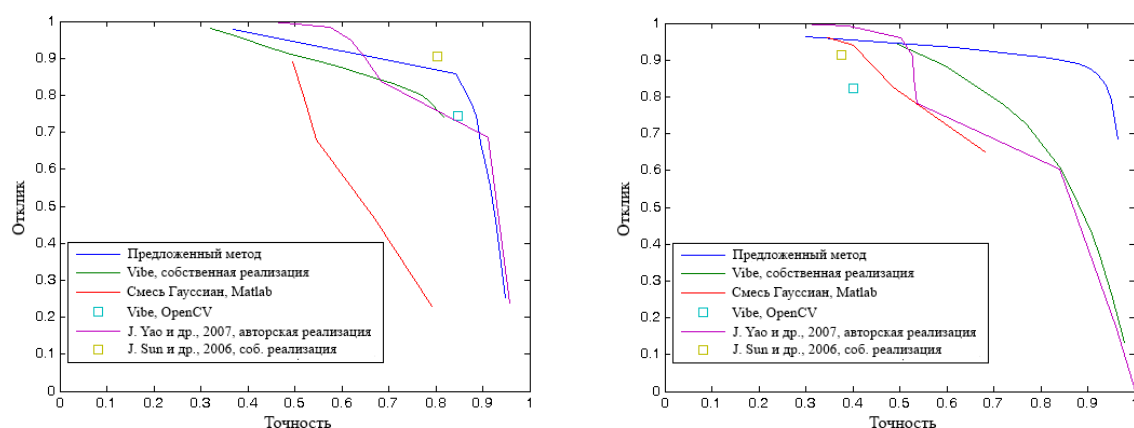


Рисунок 3. Графики точности сегментации видео на собственной выборке (слева) и на наборе видеопоследовательностей из университета Техаса (справа).

Также в рамках диссертации был предложен метод учета характерной формы объектов переднего плана для рассматриваемого сценария видеонаблюдения, который может использоваться поверх любого алгоритма вычитания фона для уточнения маски переднего плана. Метод опирается на то, что при рассматриваемом ракурсе съёмки маскам людей присущи определенные закономерности. В частности для стоящего человека, у которого ноги выходят за нижнюю границу изображения, в большинстве случаев маска почти монотонно расширяется от верхней части к нижней. Если в поле зрения камеры попадают ноги человека, тогда при просмотре изображения сверху вниз вначале маска расширяется, а потом опять сужается. На основе описанных наблюдений предлагается построить 2 мо-

⁷ J. Yao, J. M. Odobez, Multi-layer background subtraction based on color and texture., Proc. Computer Vision and Pattern Recognition, pp. 1-8, 2007

⁸ J. Sun, W. Zhang, X. Tang, and H. Y. Shum, Background cut, Proc. Europ. Conf. on Computer Vision, pp. 628-641, 2006

дели маски. Пусть $M_{x,y}$ - маска переднего плана, $M_{x,y} = 1$ для пикселей переднего плана и $M_{x,y} = 0$ для фона. Тогда:

- Модель 1. Для $\forall x \exists h(x)$: $M_{x,y} = 1$ для $y \leq h(x)$ и $M_{x,y} = 0$ для $y > h(x)$
- Модель 2. Для $\forall x \exists h_1(x)$ и $\exists h_2(x)$: $M_{x,y} = 1$ для $y \in [h_1(x), h_2(x)]$ и $M_{x,y} = 0$ для $y < h_1(x) \vee y > h_2(x)$

На рис. 4 показаны примеры кадров видео для рассматриваемого сценария видеонаблюдения, соответствующие эталонные маски переднего плана, а также видоизмененные маски, удовлетворяющие ограничениям модели 1 и 2. Видно, что маски отличаются незначительно. Идея предлагаемой модификации состоит в том, чтобы ограничить вид получаемой маски переднего плана одной из двух определенных ранее моделей. При этом получение идеально точной маски станет недостижимым, однако общая точность алгоритма благодаря дополнительным ограничениям может повыситься.

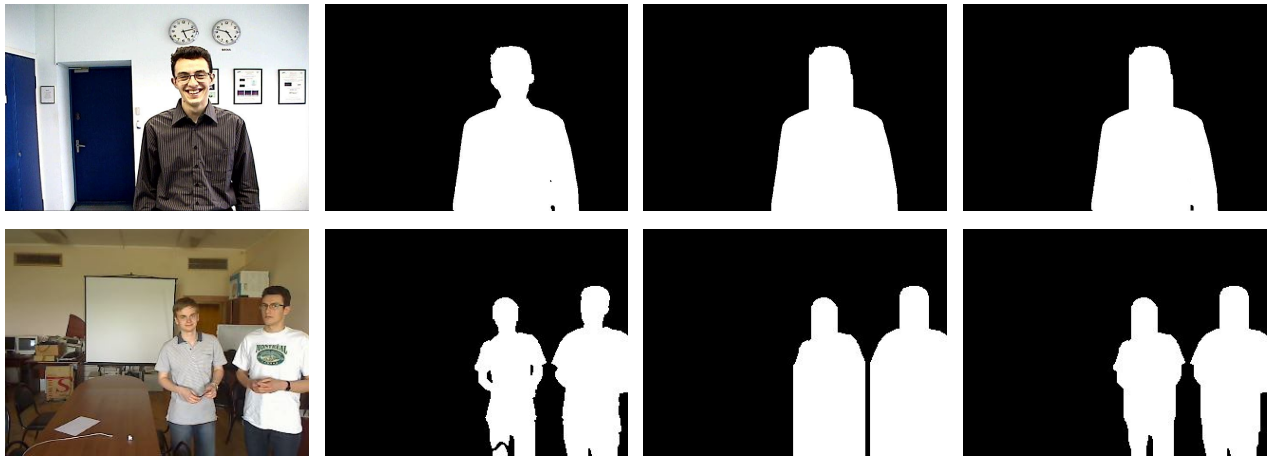


Рисунок 4. Примеры моделей масок объектов переднего плана.

Для большинства алгоритмов, работающих независимо с отдельными пикселями или блоками изображения, данная модификация сводится к пост-обработке результатов. Рассмотрим случай попиксельного алгоритма и модели маски 1. Пусть $p(I_{x,y} | M_{x,y})$ - вероятность того, что пиксель (x, y) со значением маски $M_{x,y}$ имеет цвет $I_{x,y}$. По определению модели 1 для каждого x существует $h(x)$, при котором все пиксели выше $h(x)$ отнесены к фону, а остальные пиксели отнесены к переднему плану. Т.о. значение $h(x)$ полностью определяет маску в данном столбце. Требуется максимизировать $p(h(x) = Y | I) \sim p(I_1, \dots, I_{Y-1}, I_Y, \dots, I_H | 0, \dots, 0, 1, \dots, 1)$, где H - высота кадра, а индексы x опущены для наглядности. Т.к. $I_1, \dots, I_H$ условно независимы от маски M , то $p(I_1, \dots, I_{Y-1}, I_Y, \dots, I_H | 0, \dots, 0, 1, \dots, 1) = \prod_{y < Y} p(I_y | 0) \cdot \prod_{y \geq Y} p(I_y | 1)$ Беря логарифм вероятности произведение

переходит в сумму $\ln p(h(x) = Y / I) = \sum_{y < Y} \ln p(I_y | 0) + \sum_{y \geq Y} \ln p(I_y | 1) + Const$. Если заранее вычислить интегральные суммы $F_{x,Y} = \sum_{y \geq Y} \ln p(I_{x,y} | 1)$ и $B_{x,Y} = \sum_{y < Y} \ln p(I_{x,y} | 0)$, то для максимизации вероятности $h(x)$ достаточно найти максимум по каждому столбцу $\arg \max(\ln p(h(x) = Y)) = \arg \max(F(x, Y) + B(x, Y))$

В случае с моделью маски 2 в каждом столбце нужно найти уже 2 значения $h_1(x) < h_2(x)$, и задача решается методом динамического программирования по каждому столбцу.

Также показывается, как предложенный метод можно применять для модификации алгоритмов, основанных на Марковских случайных полях, таким образом, чтобы получающиеся маски удовлетворяли моделям 1 и 2.

На рис. 5 демонстрируется повышение точности разработанного поблочного алгоритма вычитания фона, а также алгоритма Vibe при использовании предложенного метода на собственной выборке видеопоследовательностей. Важными достоинствами метода является отсутствие каких-либо дополнительных параметров, а также высокая скорость работы, благодаря которой она практически не увеличивает время работы вычитания фона.

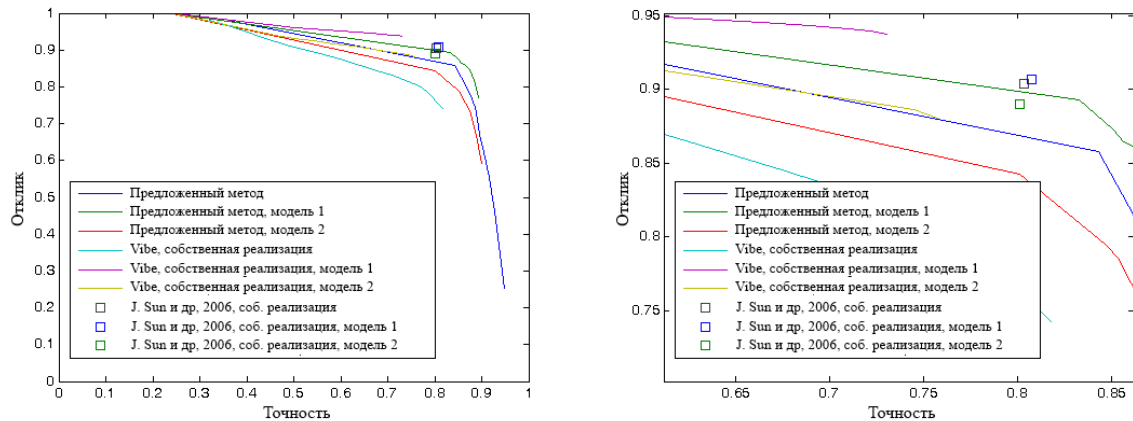


Рисунок 5. График точности сегментации видео без и с использованием разработанной модификации (справа - приближенный вариант).

Во второй главе рассматривается задача сопровождения людей в видео, т.е. отслеживания траектории их движения в области видимости камеры с течением времени. В рамках диссертации используется следующая формулировка данной задачи. На вход алгоритму последовательно подаются кадры видео $I_1, \dots, I_k$, полученные стационарной камерой видеонаблюдения. Сценарий видеонаблюдения при этом удовлетворяет условиям, описанным ранее. Для каждого кадра I_i имеется маска переднего плана M_i и набор ограничивающих прямоугольников лиц людей на этом кадре $\{Face_j\}_i$, полученные с помощью

алгоритмов вычитания фона и выделения лиц на изображении. Требуется определить положение всех людей $\{Person_k\}_i$, попавших в поле зрения камер, для каждого кадра видеопотока, а также сопоставить их с выделенными на предыдущем кадре людьми, если они на нём были.

Стоит отметить, что формулировки задачи сопровождения объектов могут отличаться. Например, обнаружение объектов иногда выносят за рамки алгоритма сопровождения, а иногда наоборот, получение маски переднего плана могут включить в задачу сопровождения. Это связано с тем, что сам процесс выделения объектов или вычитания фона может являться составной частью алгоритма сопровождения объектов.

В рамках общей задачи анализа аудитории сопровождение людей решает несколько подзадач. Во-первых, сопровождение позволяет повысить точность подсчета числа людей и их классификации, т.к. в случае взаимного перекрытия людей не всегда возможно надежно их выделить по одному кадру, а при сильных поворотах головы точность классификации будет слишком низкой. Сопровождение объектов в свою очередь позволяет выделять и классифицировать людей лишь по наиболее удачным кадрам. Во-вторых, оно позволяет снизить вычислительную нагрузку благодаря более редкой классификации людей. В-третьих, сопровождение людей «приближает» к подсчету уникальных посетителей, что для многих задач, например, маркетинговых, является более важной характеристикой. Для полноценного подсчета уникальных посетителей при анализе аудитории дополнительно нужно использовать идентификацию людей, чтобы сопоставлять людей в случае выхода и повторного захода в поле зрения камеры или при потере сопровождения.

Можно выделить следующие основные отличия рассматриваемого сценария видеонаблюдения:

1. Нижняя часть тела человека и области земли не всегда видна на изображении. Поэтому не видна точка контакта человека и поверхности земли, что затрудняет определение положения человека относительно камеры и сопровождение в системе координат, связанной с землёй. В результате сопровождение можно проводить только в системе координат изображения.
2. Люди часто перекрыты другими людьми в течение долгого интервала времени, что не так характерно для обычных сценариев видеонаблюдения с высоко расположенными камерами.
3. Изображения людей занимают большую часть кадра. Во-первых, это затрудняет поддержание модели фона. Во-вторых, это приводит к проблемам с обработкой ситуаций входа и выхода из сцены, поскольку края изображения могут быть перекрыты людьми, и новые объекты могут появиться внезапно сразу в центре изображения.

4. Нелинейное движение людей в поле зрения из-за частых внезапных остановок людей возле монитора с камерой, когда их привлекает изображение на мониторе. После остановки люди могут как продолжить движение в прежнем направлении, так и развернуться в обратную сторону. В этом сценарии затруднено предсказание положения людей в новом кадре.
5. Алгоритм должен реагировать на события с минимальной задержкой из-за своего возможного применения в информационных стендах. Это означает, что алгоритм должен обрабатывать кадры последовательно по одному из видеопотока, а не в совокупности анализировать всю видеопоследовательность.

По результатам проведенного обзора с учётом указанных ограничений для предлагаемого алгоритма выбран подход на основе фильтра частиц, который не накладывает ограничения на модели наблюдения и движения и может использоваться для обработки видеопотока в реальном времени. Для корректного разрешения коллизий при перекрытии объекта в предложенном алгоритме явно моделируется и учитывается видимость объекта. Также для различения наблюдаемых объектов используется набор классификаторов «каждый с каждым», обученных налёту и использующих цветовые и текстурные признаки объектов.

Рассмотрим подробнее метод фильтра частиц. Обозначим состояние объекта (например, человека) на кадре t через S_t , а наблюдение через Z_t . Все наблюдения до t включительно обозначим через Z^t . Тогда задача сопровождения сводится к оценке состояния $p(S_t | Z^t)$. Если мы используем модель Марковского процесса первого порядка и хотим оценивать состояние объекта «на лёту», то этот процесс может быть разбит на 2 части: предсказание состояния по предыдущему состоянию системы и последующее обновление состояния с учетом наблюдаемых данных:

$$p(S_t | Z^{t-1}) = \int p(S_t | S_{t-1}) p(S_{t-1} | Z^{t-1}) dS_{t-1}$$

$$p(S_t | Z^t) = \frac{p(Z_t | S_t) p(S_t | Z^{t-1})}{p(Z_t | Z^{t-1})}$$

Метод фильтра частиц аппроксимирует вероятностное распределение взвешенной суммой N частиц:

$$p(S) \approx \sum_{i=1}^N w^i \delta(S - s^i), \text{ где } w^i, \sum_{i=1}^N w^i = 1 - \text{веса частиц.}$$

В рамках диссертации был использован метод SIR (Sequential importance resampling). На этапе предсказания состояние частиц случайным образом изменяется на основе модели движения $p(s_t^i | s_{t-1}^i)$. После этого веса частиц изменяются на основе моде-

ли наблюдения $p(z_t^i | s_t^i)$. Затем происходит нормализация весов частиц для обеспечения $\sum_{i=1}^N w^i = 1$. Для разработки алгоритма на базе фильтра частиц нужно определить модель объекта и задать вероятности $p(s_t^i | s_{t-1}^i)$ и $p(z_t^i | s_t^i)$.

В предлагаемом алгоритме в качестве модели объекта используется четверка $\{x, wd, v, l\}$, где x – х-координата центра человека, wd – его ширина, v – скорость, а l – метка {«фон», «передний план»}. Если l = «передний план», то человек считается по большей части видимым, если l = «фон», то закрытым другим человеком. Y-координаты человека явно не моделируются – предполагается, что верхняя граница человека ограничена маской переднего плана, а нижняя граница – нижней границей изображения, т.к. обычно весь человек на кадре не виден. Важно отметить, что область значений x превышает ширину кадра на величину, равную максимальной допустимой ширине человека. Это позволяет моделировать ситуации, когда человек частично не виден на кадре.

Явное использование метки l является одним из важных отличий предлагаемого алгоритма от существующих методов. Благодаря ее наличию, алгоритм способен корректно обрабатывать ситуации перекрытия между людьми, когда на протяжении длительного времени нельзя точно сказать, за кем именно стоит отслеживаемый человек, и лишь спустя какое-то время, когда он выйдет на передний план, можно принимать решение.

С учетом использования фильтра частиц, объект в каждый момент времени описывается набором частиц $s_k^i = \{x_k^i, wd_k^i, v_k^i, l_k^i\}$. Если сумма весов всех частиц с меткой l_k^i = «фон» больше суммы весов частиц с меткой l_k^i = «передний план», то считается, что человек стоит на заднем плане, иначе – на переднем.

Общая схема работы алгоритма показана на рис. 6.

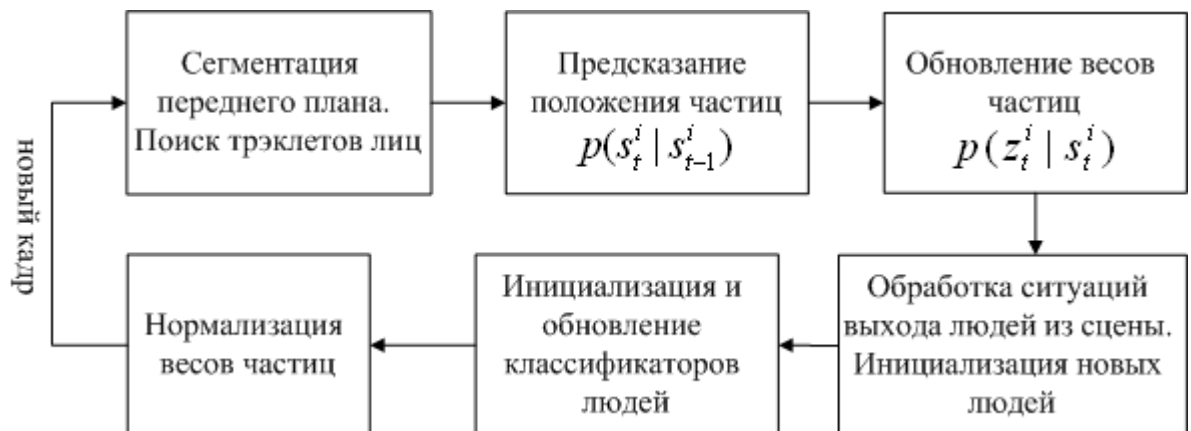


Рисунок 6. Схема работы алгоритма сопровождения людей в видео.

На вход алгоритма вместе с очередным кадром видео подаётся маска переднего плана и набор найденных лиц, которые сопоставляются с лицами с предыдущих кадров, образуя треклеты лиц.

Затем происходит предсказание смещения частиц на основе модели движения объектов по следующим правилам:

$$x_t = x_{t-1} + v_{t-1} + N(0, \sigma_x)$$

$$wd_t = wd_{t-1} + N(0, \sigma_{wd})$$

$$v_t = v_{t-1} * (1 - \beta) + \beta * (x_t - x_{t-1}) + N(0, \sigma_v)$$

где $N(\mu, \sigma)$ - нормальное распределение с математическим ожиданием μ и дисперсией σ . При этом, если предсказанное положение частицы выходит за границы связной компоненты, то ее положение вычисляется повторно.

Изменение метки моделируется распределением Бернулли:

$$p(l_t^i | s_{t-1}^i) = \alpha \cdot l_{t-1}^i + (1 - \alpha)(1 - l_{t-1}^i), \text{ где } \alpha - \text{параметр.}$$

Т.о. каждый кадр есть константная вероятность изменения метки на противоположную.

Далее происходит обновление весов частиц. Большой вес получают те частицы, которые лучше согласуются с текущим кадром (наблюдением). Для оценки согласования частицы с изображением используется набор бинарных классификаторов «каждый с каждым» на основе бустинга, обучаемых «на лету». Использование схемы «каждый с каждым» обеспечивает большую точность по сравнению со схемой «один против всех», т.к. в обучающую выборку включаются в качестве отрицательных примеров только примеры одного объекта, а распределение признаков одного объекта моделировать проще, чем нескольких одновременно. В качестве признаков используются цветовая гистограмма в цветовом пространстве RGB и гистограмма однородных локальных бинарных шаблонов (Local Binary Patterns, LBP). При оценке правдоподобия наблюдения конкретного человека берутся выходы классификаторов, обученных на нём и ближайших людях, и вычисляется минимум по ним:

$$p(z_t | s_t^i) = \begin{cases} \min_{j \in Blob(tr), tr \neq j} BinaryClassifier_{(j, tr)}(z_t, s_t^i) & l_t^i = true \\ 0.5 & l_t^i = false \end{cases}$$

Если человек в данный момент перекрыт или находится вне изображения, то правдоподобие наблюдения устанавливается равным 0.5.

После этого выполняется анализ, вышел ли какой-либо человек из сцены, или наоборот, появился новый человек. Выход человека определяется при одновременном выполнении нескольких условий:

- суммарный вес частиц вовне изображения должен быть больше порога
- частицы, находящиеся внутри изображения, преимущественно фоновые
- с тех пор, как какая-то часть частиц человека оказалась вовне изображения, прошло не менее K кадров.

С учетом того, что люди занимают большую часть кадра, регулярно возникают ситуации, когда одновременно один человек выходит из кадра, а другой входит в той же части кадра. В этом случае частицы уходящего человека постепенно перескакивают на входящего, а классификатор человека успевает переобучиться, что приводит к перескоку сопровождения. Для решения данной проблемы запоминаются признаки человека за последние L кадров. И если в подобной ситуации внешний вид (признаки) человека резко меняется, то значит нужно инициализировать нового человека.

Также инициализация нового человека происходит в случае возникновения большой связной области в маске переднего плана, в которой не находится ни один из отслеживаемых людей. Другим вариантом является обнаружение лица, которое нельзя сопоставить текущим людям.

При инициализации нового человека создаются набор классификаторов для сравнения данного человека с остальными, и обновляются классификаторы для всех остальных людей на основе новых данных. Под конец происходит нормализация весов частиц.

Для экспериментальной оценки предложенного алгоритма была использована собственная выборка, поскольку для рассматриваемого сценария видеонаблюдения доступных тестовых баз найдено не было. Выборка состоит из 35 видеопоследовательностей, содержащих от 2 до 13 людей.

В качестве базового алгоритма для сравнения использовался метод фильтра частиц с классификаторами вида «один-против-всех» и без метки видимости. Базовый алгоритм близок к алгоритму из работы Бернштайна и др.⁹, но адаптирован под рассматриваемый сценарий видеонаблюдения. Далее поверх него по очереди добавлялись элементы предложенного алгоритма: метка видимости (МВ), классификатор вида «каждый с каждым» (КК) и сохранение вида объекта в течение L кадров для разрешений коллизий одновременного входа-выхода людей (ВВ). В таблице 1 описаны используемые в экспериментах метрики точности. В таблице 2 показаны результаты экспериментов. Видно, что все 3 предложенные модификации последовательно повышают точность алгоритма. Более де-

⁹ M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, Luc Van Gool, Robust Tracking-by-Detection using a Detector Confidence Particle Filter, Proc. International Conference on Computer Vision, pp. 1515-1522, 2009

тальный анализ результатов показывает, что уменьшение ошибок в основном происходит лишь в сложных случаях, при длительных перекрытиях между людьми. На рис. 7 приведен пример работы алгоритма.

Название	Определение
Перескок метки	Число раз, когда отслеживаемая траектория меняет соответствующую ей эталонную метку
Фрагменты	Число раз, когда эталонная траектория прерывается в результатах сопровождения
Отклик	(покадрово) число правильно сопоставленных объектов / общее число эталонных объектов
Точность	(покадрово) число правильно сопоставленных объектов / общее число найденных объектов

Таблица 1. Используемые метрики при тестировании алгоритма сопровождения объектов

Алгоритм	Перескок метки	Фрагменты	Отклик	Точность
Базовый	25	20	86%	87%
Базовый + MB	17	12	87%	89%
Базовый + MB + КК	16	5	90%	90%
Базовый + MB + КК + ВВ	10	4	90%	91%

Таблица 2. Результаты тестирования алгоритма сопровождения людей в видео



Рисунок 7. Пример работы алгоритма сопровождения людей. Под кадрами показано текущее распределение частиц объекта. Область над линией показывает видимые частицы, область под линией – частицы, которые в данный момент считаются перекрытыми.

В третьей главе рассматривается задача классификации человека по изображению его лица. На входе дано изображение I и положение антропометрических точек лица человека¹⁰. Требуется получить значение l выбранного атрибута человека L . Примерами возможных атрибутов являются пол, этническая принадлежность, возраст, наличие очков на лице и пр.

Часть атрибутов (пол, наличие очков, факт открытого рта и пр.) являются бинарными. Часть атрибутов (этническая принадлежность) являются перечислимым типом, однако их классификация сводится к нескольким бинарным задачам. Оценка возраста во многих работах сводится к классификации на несколько возрастных интервалов. Поэтому в данной работе рассматриваются только бинарные атрибуты $L = \{-1, 1\}$.

Классификация человека по изображению лица может использоваться для аннотации изображений и поиска в архивах. В рамках задачи анализа аудитории классификация людей может применяться для подсчета статистики (например, просмотров рекламы), для направленной рекламы (для заданной половозрастной группы) и для интеллектуальных человеко-машинных интерфейсов.

Большинство существующих алгоритмов классификации людей по лицу используют одинаковую схему работы: нормализация фотографии, вычисление признаков и непосредственно классификация на основе алгоритма машинного обучения. Разные методы отличаются лишь конкретными реализациями отдельных этапов. Для нормализации лица чаще всего используется комбинация поворота лица, при котором отрезок между глазами человека становится горизонтальным, масштабирования и обрезки изображения. В качестве признаков в разных статьях предлагалось использовать локальные бинарные шаблоны, биологически обусловленные признаки, признаки Хаара и пр. Для классификации используются бустинг, метод опорных векторов, комитет решающих деревьев, нейросети.

В рамках диссертации предлагается новая схема классификации людей по изображению лица на основе так называемых сравнительных признаков внешности (simile classifiers), которые показывают сходство между чертами лица на тестовом изображении с чертами лица какого-то известного человека. Например, насколько губы данной женщины похожи на губы актрисы Анжелины Джоли (или наоборот, совсем не похожи). Идею подхода можно описать следующим образом: если губы, нос и глаза человека больше похожи

¹⁰ Для выделения антропометрических точек лица человека на изображении использовался алгоритм D. Cristinacce, T. Cootes, Boosted Regression Active Shape Models, Proc. British Machine Vision Conference, 2007

на соответствующие черты женщин, чем мужчин, то, скорее всего, данный человек – женщина. Данная схема проиллюстрирована на рис. 8.

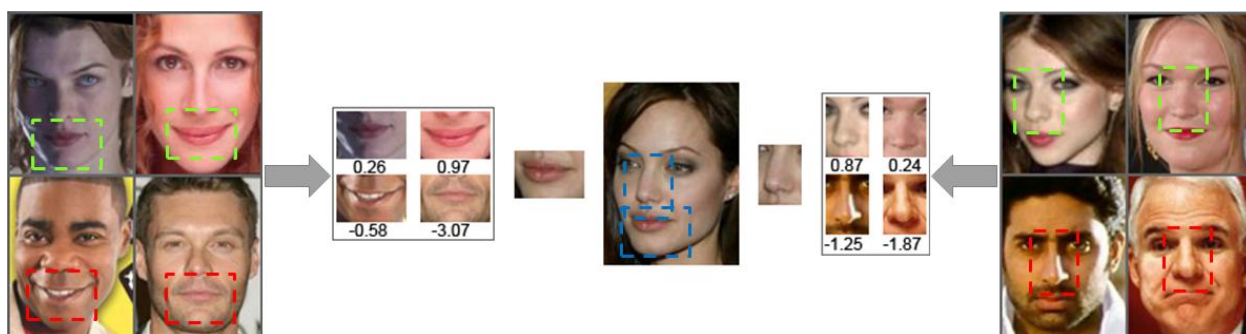


Рисунок 8. Иллюстрация схемы работы предложенного алгоритма классификации человека по изображению лица.

Рассмотрим более подробно предлагаемый алгоритм.

Для построения алгоритма необходимо использовать выборку изображений людей, в которой для ряда человек есть множество фотографий. Этих людей можно назвать «эталонными», именно с их изображениями будет сравниваться тестовое для определения нужного атрибута.

На первом этапе происходит вычисление сравнительных признаков внешности. Изображение лица разбивается на несколько областей (компонент лица), в каждом пикселе которых вычисляются такие признаки, как цвет, величина и градиент. Далее признаки используются как напрямую, так и через объединение в гистограммы, или подсчет среднего и дисперсии. Для каждого вида признаков, каждой компоненты и каждого эталонного человека обучается отдельный классификатор, который оценивает похожесть данной компоненты на данного человека. В качестве положительных примеров обучающей выборки используются соответствующие компоненты лица, взятые из разных изображений этого человека, а в качестве отрицательных примеров используется эта же компонента, но взятая из произвольных изображений других людей. В качестве классификатора используется метод опорных векторов с ядром RBF.

На втором этапе выполняется классификация тестового лица по вектор-признаку, составленному из набора сравнительных признаков. Имеется дескриптор F размерности $N \cdot M \cdot K$, где N - количество эталонных людей, M - количество видов признаков, а K - количество компонент. Каждый элемент дескриптора $F_{n,m,k}$ является оценкой похожести k -ой компоненты тестового лица и лица n -го эталонного человека, используя m -й вид признака, и по сути является выходом соответствующего классификатора. Требуется найти класс $L(F)$ тестового человека.

С учетом высокой размерности вектора признаков относительно размера обучающей выборки, которая может приводить к переобучению классификатора, было предло-

жено использовать двухуровневый классификатор. На первом уровне K классификаторов принимают решение на основании одной компоненты:

$$L_k(F) = \text{Classifier1}_k \{F_{i,j,k}, i \in \{1, \dots, N\}, j \in \{1, \dots, M\}\}$$

На втором уровне единственный классификатор объединяет K выходов классификаторов первого уровня для итогового решения:

$$L(F) = \text{Classifier2} \{L_k(F), k \in \{1, \dots, K\}\}$$

В обоих случаях в качестве классификатора используется метод опорных векторов с ядром RBF. Стоит отметить, что при этом даже не требуется информация о метках эталонных людей. Классификаторы автоматически обучаются зависимостям вида «похожесть на данного человека означает, что тестовый человек скорее имеет метку -1, а сходство с этим человека, что метку 1».

Предложенный алгоритм был протестирован на примере задачи классификации пола человека, которая является одной из наиболее широко применяемых и исследуемых. В качестве тестовой и эталонной были выбраны базы изображений LFW и PubFig¹¹ соответственно. База LFW состоит из 13233 изображений 5749 разных человек, в то время как PubFig содержит 60000 изображений 200 разных человек. Все изображения собраны автоматически из Интернета, все лица находятся с помощью алгоритма Viola-Jones. База LFW является стандартной для сравнения алгоритмов идентификации человека по лицу. Примеры изображений из баз LFW и PubFig приведены на рис. 9. В экспериментах в качестве списка эталонных людей был взят «набор разработчика» (Development Set) базы PubFig (всего 59 людей), использовалась 5-кратная кросс-валидация на базе LFW.



Рисунок 9. Примеры изображений из баз LFW (слева) и PubFig (справа).

Точность предложенного алгоритма составила 92.96%. Для сравнения можно использовать следующие результаты других алгоритмов:

1. В работе «Attribute and simile classifiers for face verification» классификация пола использовалась для идентификации людей по фотографии. Точные выборки, используемые для обучения и тестирования, приведены не были, но они являлись частью базы PubFig и LFW. Была заявлена точность в 81.22%.

¹¹ N. Kumar, A. C. Berg, P. N. Belhumeur, S. K. Nayar, Attribute and simile classifiers for face verification, Proc. International Conference on Computer Vision, 2009.

2. В статье «Learning local binary patterns for gender classification on real-world face images» для экспериментов также использовалась 5-кратная кросс-валидация на базе LFW. При этом из тестирования были исключены сложные для обработки фотографии, на которых лица запечатлены под значительным углом или для которых оказалось сложно вручную разметить пол. В результате была показана точность в 94.81% при использовании 7443 изображения из 13233 имеющихся.
3. Собственная реализация алгоритма Гуо и др.¹² показала точность в 90.18%.

Экспериментальное сравнение показало, что предложенный метод значительно превосходит по точности классификации первую работу, в которой используются те же признаки, что и предложенном методе. При этом достигнутая точность немного уступает второй работе, в которой были исключены из рассмотрения чуть менее половины примеров изображений, являющихся наиболее сложными для классификации, а также использовались более сложные признаки классификаторов. Также показанная точность превосходит точность третьей работы, при этом используя более простые признаки классификаторов.

В заключении перечислены основные результаты работы.

Основные результаты работы

В результате работы был предложен комплекс новых алгоритмов для анализа аудитории по данным видеонаблюдения. Рассматривается сценарий видеонаблюдения, в котором камера закреплена рядом с демонстрационным экраном на высоте 1-2 метра от земли и направлена на людей, подходящих к монитору. Для данного сценария видеонаблюдения предложенные алгоритмы превосходят по точности существующие аналоги.

1. Разработан новый алгоритм вычитания фона на основе поблочных классификаторов. Алгоритм показывает высокую устойчивость к изменению освещения, превосходя существующие аналоги по точности.
2. Предложен метод учета характерной формы изображений людей для модификации произвольных алгоритмов вычитания фона для рассматриваемого сценария видеонаблюдения. Метод повышает точность сегментации видео без заметного повышения времени обработки.

¹² G. Guo, G. Mu, Y. Fu, and T. S. Huang, Human age estimation using bio-inspired features, Proc. CVPR, 2009

3. Предложен новый алгоритм сопровождения людей в видео на основе метода фильтра частиц. По сравнению с базовым алгоритмом предложенный алгоритм более устойчив к перекрытиям и продемонстрировал более высокую точность слежения.
4. Предложена новая схема алгоритма классификации людей по изображению лица. На примере задачи классификации пола человека предложенный алгоритм превосходит по точности существующие аналоги.

Публикации по теме диссертации

1. Konushin V., Konushin A. Improvement of background subtraction by mask constraints // Proceedings of GraphiCon. 2010. P. 96-99.
2. Шальнов Е., Кононов В., Конушин В. Алгоритм вычитания фона, основанный на поблочных классификаторах // Труды конференции ГрафиКон. 2011. С. 227-230.
3. Конушин В., Кононов В., Конушин А. Алгоритм слежения за людьми для построения интеллектуального интерфейса человек-компьютер на основе фильтра частиц // Программирование. 2011. № 5. С. 27-38.
4. Kononov V., Konushin V., Konushin A. People Tracking Algorithm for Human Height Mounted Cameras // Springer Lecture Notes in Computer Science, Pattern Recognition. 2011. V. 6835. P. 163-172.
5. Konushin V., Lukina T., Kuharenko A., Konushin A. Simile Classifiers for Face Classification // Proceedings of GraphiCon. 2012. P. 108-112.
6. Конушин В., Лукина Т., Кухаренко А., Конушин А. Классификация людей по изображению лица на основе сравнительных признаков внешности // Системы и средства информатики. 2013. Т. 23. №. 2. С. 34-45.

Личный вклад автора

Все результаты, составляющие основное содержание диссертации, получены автором самостоятельно. В работе [1] автором предложен метод учета характерной формы человека в рассматриваемом сценарии видеонаблюдения для модификации методов вычитания фона. В работе [2] автором предложена идея использования быстрых классификаторов на основе алгоритма бустинг и простых признаков для классификации блоков кадров видео на фон и передний план. В работах [3, 4] автором предложен алгоритм сопровождения людей в видео на основе фильтра частиц для рассматриваемого сценария, с явным моделированием состояния видимости объекта и использованием классификаторов «каждый-с-каждым» для различения людей. В работах [5, 6] автору принадлежит идея классификации атрибутов человека на основе классификации сравнительных признаков внешности.