Алгоритм распознавания людей в видеопоследовательности на основе случайных патчей

Вадим Конушин, Глеб Кривовязь, Антон Конушин
Лаборатория Компьютерной Графики и Мультимедиа, МГУ им. М.В.Ломоносова, Москва, Россия
{vadim, gleb.krivovyaz, ktosh}@ graphics.cs.msu.ru

Аннотация

Распознавание людей по видео является широко востребованной задачей. Из-за того, что алгоритмы распознавания людей по лицу до сих пор не дают приемлемого качества распознавания в случае низкого разрешения изображения/видео, а также произвольного освещения, в последнее время всё большее распространение находит распознавание людей, учитывающее одежду человека.

В данной статье предложен новый алгоритм распознавания человека по видео, основанный на извлечении из видео большого количества случайных патчей. В отличие от большинства современных алгоритмов, предложенный подход не опирается на маску, полученную с помощью методов вычитания фона, из-за чего он способен работать на видео с произвольным сложным фоном.

Ключевые слова: *Распознавание человека по видео, машинное обучение*

1. ВВЕДЕНИЕ

Распознавание людей является одной из самых бурно развивающихся областей компьютерного зрения. Во многом это связано с тем, что данная область имеет большое количетсво применений на практике: в охранных системах, для рекламы и пр.

Одними из самых надеждых считаются методы распознавания по отпечаткам пальцев или по радужке глаза. Однако эти методы являются инвазивными – т.е. для распознавания человека требуется его «сотрудничество» - например, положить палец в устройство по считыванию отпечатков. Поэтому, представляет интерес разработка алгоритма, способного распознавать людей лишь по их фотографии/видео.

Наиболее исследованными на данный момент являются алгоритмы распознавания по лицу [11]. Но, несмотря на весь прогресс в этой области, для качественного распознавания по лицу по-прежнему требуется большое разрешение изображения, а также контролируемое освещение. Вдобавок, человек зачастую может просто не посмотреть в камеру.

Из-за всех этих недостатков, в последнее время всё чаще используется дополнительная информация: об одежде, о цвете волос и пр. Эта информация не является инвариантной для человека – он может перекрасить волосы, переодеть одежду. Однако на протяжении небольшого промежутка времени (например, один день), все эти признаки часто остаются неизменными. В данной статье предлагается новый алгоритм распознавания человека по видео, снятому со стационарной камеры.

Рис 1. Пример роликов из созданной тестовой выборки.

Данный алгоритм, мотивированный статьей [5], использует случайный набор патчей, взятых из произвольных кадров видео, и классифицирует их с помощью случайного леса деревьев. Его основным достоинством является то, что он не опирается на маску переднего плана, полученную с помощью алгоритмов вычитания фона, а значит, допускает произвольный, сложный фон, с которым не справляются современные алгоритмы вычитания фона.

В виду отсутствия публично-доступных подходящих баз видеороликов, для тестирования предложенного алгоритма была создана собственная выборка видеороликов. Примеры кадров из этих видеороликов показаны на Рис. 1.

Оставшаяся часть статьи организована следующим образом: в секции 2 приведен обзор существующих методов, в секции 3 более подробно описана собранная база видеороликов, секция 4 описывает предложенный алгоритм, в секции 5 рассказывается о проведенных экспериментах, и, наконец, в секции 6 даётся заключение статьи.

2. СУЩЕСТВУЮЩИЕ ПОДХОДЫ

В основном распознавание людей, учитывающее одежду, используется в 2 задачах: аннотации изобрежений/видео и в системах видеонаблюдения.

В алгоритмах аннотации изображений/видео [1], [2] одной из основных проблем является сегментация человека. Современные алгоритмы сегментации не позволяют автоматически получить точную маску объекта на произвольном фоне. Поэтому стандартной схемой является следующая: вначале на изображении находится лицо человека, после чего в качестве области одежды берется прямоугольник под лицом человека. Координаты и размер прямоугольника задаются эвристически.

Понятно, что такой прямоугольник содержит лишь небольшую часть от всей области одежды. Более того, в некоторых случаях, например, если человек стоит нефронтально по отношению к камере, этот прямоугольник может частично захватить область фона.

В [3] используется более сложный алгоритм сегментации, однако ему тоже для первого приближения необходимо найти

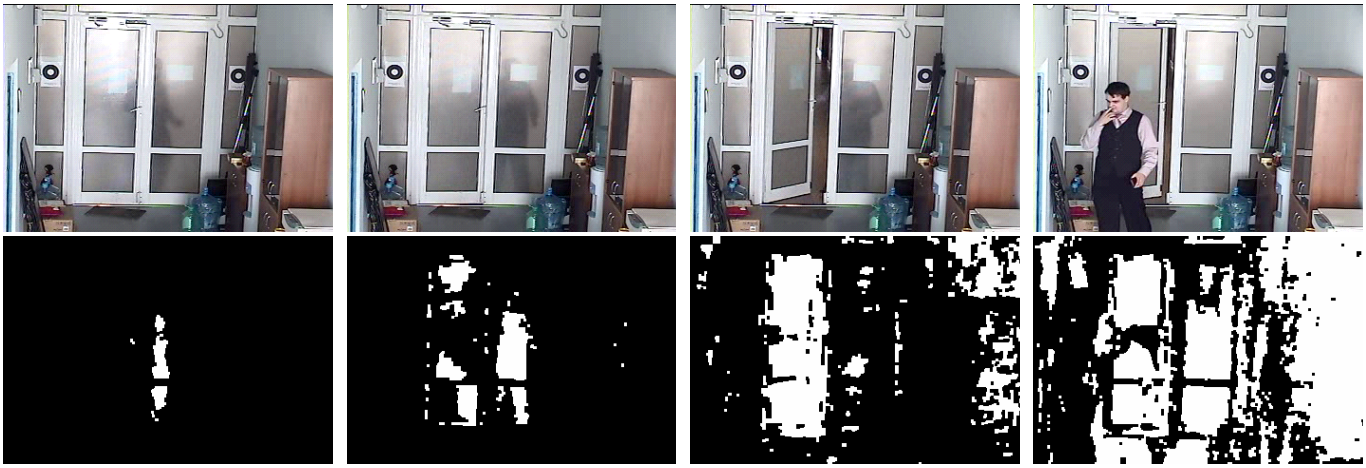

Рис 2. Демонстрация работы алгоритма вычитания фона.

лицо человека. С учетом того, что современные алгоритмы нахождения лица могут относительно надежно находить лишь фронтальные лица, то, в случаях, когда человек ни разу не посмотрит прямо в камеру (или его лицо будет чем-то загорожено), данный подход не сработает.

В [7] предложен метод, который на серии фотографий, используя информацию об одежде, цвете кожи и волос человека, позволяет обнаружить лица на тех фотографиях, на которых алгоритм нахождения лиц ничего не нашел. Однако, для получения цветовой модели одежды, кожи и волос, необходимо, чтобы лицо нашлось хотя бы на части из изображений.

В системах видеонаблюдения [6], [10] маска человека находится с помощью алгоритмов вычитания фона. И основной упор в этих статьях уже делается на том, что лучше всего использовать для описания одежды, и как ее лучше классифицировать.

Однако современные алгоритмы вычитания фона хорошо работают лишь в не очень сложных случаях. Например, чаще всего камера наблюдает за хорошо освещенным коридором. Но, как мы обнаружили на нашей выборке видеороликов, в более сложных случаях, эти алгоритмы могут давать неудовлетворительную маску. Более подробно об этом описывается в секции 3.

Предлагаемый в данной статье метод во многом основывается на статье [5]. Авторы [5] предложили метод для классификации изображений, который, в частности, тестировали на задачах классификации различных домов, разных типов сцен. В предложенном алгоритме, мы применяем похожую схему для классификации людей на видео. Главным отличием является то, что если в [5] классифицируемый объект занимал либо всё, либо большую часть изображения, то в нашем случае, человек занимает лишь маленькую часть от общего пространственно-временного объема видео.

3. ИСПОЛЬЗУЕМАЯ ВЫБОРКА

В рамках данной работы, нами была собрана собственная выборка видеороликов. Эти видеоролики записывались камерой, висящей под потолком, и направленной на входную дверь в лабораторию. Запись проводилась в течении 8 дней, всего было собрано 463 видеоролика. Разрешение видео – 352 на 240, средняя длительность ролика – около 10 секунд. Примеры кадров из собранных видеороликов показаны на Рис. 1. Всего на разных видеороликах присутствует 25 разных людей.

Каждый ролик был вручную аннотирован на предмет присутствующих на нем людей (их меток), а также флагом – входит данный человек в комнату или наоборот выходит из нее.

Во многих случаях было невозможно разбить общее видео на несколько роликов, в каждом из которых присутствовал бы лишь один человек – например, когда двое-трое человек одновременно выходят из комнаты. В данной работе такие ролики были удалены из рассмотрения. Стоит отметить, что существующие методы тоже пока не способны обрабатывать такие случаи (надежно сегментировать людей друг от друга).

С учетом выкидывания таких роликов, в каждый конкретный день людей, появляющихся на двух и более роликах в среднем было порядка 5-8 человек.

Т.к. изначально планировалось использовать стандартную схему по распознаванию людей, нами были протестированы несколько алгоритмов вычитания фона [8],[9] на собранных данных. Оказалось, что используемые видеоролики слишком сложны для испытываемых алгоритмов – входящие/выходящие люди очень нестабильно сегментировались от фона. Один из наиболее плохих примеров работы алгоритмов вычитания фона продемонстрирован на Рис. 2.

Основными причинами таких результатов являются:

- Открывающаяся дверь (и абсолютно произвольный фон за ней)
- Полупрозрачное стекло, за которым видны тени людей, стоящих за стенкой
- Нестабильное освещение.

При подгонке параметров алгоритмов на одних данных, алгоритм вычитания фона давал плохие результаты на других примерах. Общих параметров, подходящих для всех роликов найти не удалось.

В любом случае, даже если путем дополнительных усилий и удалось бы подобрать наилучшие параметры, добавить какие-нибудь эвристические правила (учитывающие

местонахождение двери на кадрах), полученный алгоритм работал бы лишь для данной сцены. Для запуска его на других сценах, пришлось бы заново искать параметры и эвристики. Для автоматического подбора параметров, для каждой новой сцены необходимо ручными (полуавтоматическими) методами отсегментировать большое количество роликов, что занимает существенное время. В нашем же случае, мы ограничиваемся лишь метками (идентификаторами) людей, присутствующих на видео.

Дополнительно, про собранную базу видеороликов стоит отметить, что в большинстве случаев люди меняли одежду каждый день. И даже на протяжении одного одня, один и тот же человек мог один раз появиться как в куртке, так и без нее.

4. ПРЕДЛОЖЕННЫЙ АЛГОРИТМ

4.1 Извлекаемые патчи

Вначале, из тренировочных видеороликов извлекается большое число (N_{Train}) квадратных патчей. Номер кадра, размер и положение каждого кадра выбирается случайным образом. После этого все извлеченные патчи масштабируются до фиксированного размера *rxr* и переводятся в цветовое пространство HSV.

Далее эти патчи вытягиваются в вектор длины $r * r * 3$ и каждому патчу присваивается метка – идентификатор человека с видеоролика, из которого взят данный патч.

Т.к. видеоролики получены с одной и той же стационарной камеры, то, в отличие от [5], становится возможным использовать в качестве признаков патча еще и его пространственное положение – координаты левого-верхнего угла (x, y) и ширину w.

Т.о. мы получаем тренировочную выборку длиной N_{Train} из векторов длиной $r * r * 3 + 3$.

4.2 Обучение и распознавание

Полученную тренировочную выборку можно подать на вход произвольному алгоритму машинному обучения.

С учетом размера входной выборки, а самое главное – с учетом длины вектора признаков каждого элемента выборки, алгоритм машинного обучения должен быть максимально быстрым.

В данном алгоритме используется случайный лес деревьев [4]. Для тренировки каждого отдельного дерева выбирается случайная подвыборка от общей тренировочной выборки. Используемые в узлах дерева функции – сравнение одной из координат входного вектора с порогом. При этом для скорости, выбор координаты и самого порога при обучении происходят абсолютно случайно. Каждое дерево строится до тех пор, пока оно не будет полностью правильно классифицировать свою тренировочную подвыборку. На этапе распознавания из тестового видеоролика также случайным образом выбирается N_{Test} квадратных патчей, после чего они также преобразуются в вектора $r * r * 3 + 3$.

Затем эти патчи подаются для классификации натренированному лесу деревьев. В итоге мы получаем $N_{Test} * K$ разных меток, где K - количество деревьев в лесе.

После этого, в качестве вероятности каждой метки L можно брать относительное число раз, когда патчи из тестового видео были классифицированы как L:

$$p(L) = \frac{\sum_{k=1}^{K} \sum_{x_i \in X} l(T_k(x_i) = L)}{N_{Test} * K},$$

где X - множество из N_{Test} векторов, $T_i(x)$ - результат классификации вектора x i-ым деревом.

Основная проблема заключается в том, что большинство патчей окажутся взятыми из фона, и не будут пересекаться с изображением человека. Поэтому для того, чтобы результат итогового распознавания оказался правильным, необходимо, чтобы извлекаемое число патчей было достаточно большим. В таком случае, патчи, соответствующие фону будут классифицироваться произвольно (и в среднем давать одинаковую вероятность всем меткам), а уже те патчи, которые пересекаются с изображением человека, будут давать бóльшую вероятность правильной метке.

5. ЭКСПЕРИМЕНТЫ

При проведении экспериментов, использовались следующие параметры:

- Число патчей в тренировочной выборке - $N_{Train} = 100000$
- Размер, к которому мастшабируются все патчи - $r = 16$
- Число патчей, извлекаемых из тестового видео - $N_{Test} = 3000$
- Число деревьев $K = 20$

Для тестирования использовалось 2 сценария:

- В первом сценарии используются лишь те ролики, на которых человек входит в комнату. В качестве тренировочных данных использовались первые 2 видеоролика каждого человека, остальные – попадали в тестовую выборку.
- Во втором сценарии уже использовались все видеоролики. Тренировочными были первые 3 ролика каждого человека, все остальные тестовые.

В обоих случаях, человек из тестовой выборки присутствовал в тренировочных данных.

Результаты работы алгоритма представлены на Рис. 3. Демонстрируемые результаты являются суммой результатов за все 8 дней наблюдения.

В качестве метрики качества используется метрика «N лучших» (TopN), которая для каждого конкретного n показывает, для какого процента всех роликов, правильная метка была среди первых n результатов.

Т.к. в один день в тренировочной выборке присутствовало максимум 14 человек, то значение метрики при $n \geq 14$ равно 100%.

Для ориентира на графиках приведен результат распознавания в случае, если бы классификация осуществлялась случайно (подбрасыванием монетки).

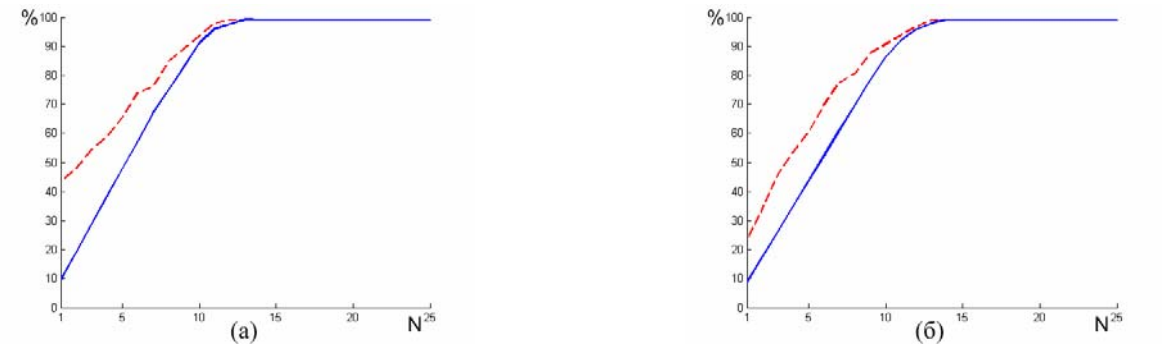

Рис 3. Результаты работы алгоритма. Красным (штриховая линия) отмечен результат работы предложенного алгоритма, синим (сплошная линия) – средний результат случайной классификации.

Из графиков видно, что полученные результаты еще далеки до того, чтобы их можно было надежно использовать в реальной системе. Например, в первом сценарии процент правильно распознанных роликов (т.е. TopN(1)) составляет лишь 45%.

Однако, также стоит учитывать и сложность входных данных. В частности, человек в один и тот же день может появляться на видео как в куртке, так и без нее. А значит, если, скажем, на первых 2 (тренировочных) роликах он будет в куртке, то распознать его на остальных роликах по используемым признакам будет практически невозможно.

Большую проблему составляет освещение. В тех случаях, когда к концу дня, когда уже темно, еще не включили свет в комнате, даже человек зачастую не сразу распознает людей на видео.

Во втором сценарии используется тот факт, что во многих случаях цвет и текстура одежды на спине очень похожа на одежду спереди. А значит, есть надежда, что алгоритм, натренированный на видео, на котором человек входит, сможет распознавать его же, но когда он выходит, и наоборот. Как видно, пока, к сожалению, при данном сценарии алгоритм сработал значительно хуже, чем в первом.

Тем не менее, полученные результаты позволяют надеяться на то, что при должной доработке данного алгоритма, он сможет достичь более хорошего процента правильного распознавания.

6. ЗАКЛЮЧЕНИЕ

В данной статье был предложен новый алгоритм распознавания человека по видео. Основным его преимуществом является то, что он не опирается на маску переднего плана, полученную с помощью алгоритмов вычитания фона, как это делают большинство существующих алгоритмов. Благодаря этому, для его тренировки достаточно предоставить лишь выборку видеороликов с меткой – идентификатором присутствующего на видео человека.

Данный алгоритм был протестирован на собственной выборке видеороликов. Несмотря на то, что полученные результаты еще далеки от идеальных, они позволяют говорить о потенциале данного подхода.

Работа выполнена при поддержке гранта РФФИ 09-01-92474-МНКС_а.

7. ЛИТЕРАТУРА

[1] Anguelov D., Lee K., Göktürk S.B., Sumengen B. *Contextual identity recognition in personal photo albums* Proc. of CVPR, pp. 1-7, 2007

[2] Everingham M. R., Sivic J., Zisserman A. *'hello! my name is... buffy' - automatic naming of characters in tv video* Proc. of BMVC, pp. 889–908, 2006

[3] Gallagher A., Chen T. *Clothing cosegmentation for recognizing people* Proc. of CVPR, No. 1, pp. 1-8, 2008

[4] Geurts P., Ernst D., Wehenkel L. *Extremely randomized trees* Machile Learning Journal, 63(1), 2006

[5] Maree R., Geurts P., Piater J., Wehenkel L., *Random Subwindows for Robust Image Classification* Proc. of CVPR, Vol. 1, pp. 34–40, 2005

[6] Nakajima C., Pontil M., Heisele B., Poggio T, *Full body person recognition system* Pattern Recognition, 2003

[7] Sivic J., Zitnick C. L., Szeliski R. *Finding People in Repeated Shots of the Same Scene* Proc. of BMVC, 2006

[8] Stauffer C., Grimson W.E.L., *Adaptive background mixture modelsfor real-time tracking* Proc. of CVPR, pp. 246-252, 1999

[9] Wren C., Azarbayejani A., Darrell T., Pentland A. *Pfinder:Real-time Tracking of the Human Body* IEEE Trans. on PAMI, Vol. 19, No. 7, pp. 780-785, 1997

[10] Yoon K., Harwood D., Davis L., *Appearance-Based Person Recognition Using Color/Path-Length Profile* Elsevier Real-Time Imaging, 2005

[11] Zhao W., Chellappa R., Phillips P. J., Rosenfeld A. *Face recognition: A literature survey*. ACM Computing Surveys, 2003

ABSTRACT

Video-based human recognition is a highly demanded task. Current face recognition algorithms still do not provide acceptable recognition rates in case of low-resolution video or in case of arbitrary lighting, hence algorithms, which take human clothes into account, are getting more and more attention.

In this paper, a new video-based human recognition algorithm is proposed, which is based on extraction of a large number of random patches. Unlike most modern algorithm, the proposed approach isn't based on a mask, obtained from background subtraction methods, and therefor it is capable of working on video with arbitrary complex background.