УДК 528.854

ОПТИМИЗИРОВАННЫЙ МЕТОД РЕКОНСТРУКЦИИ ПРОСТРАНСТВЕННОЙ КОНФИГУРАЦИИ ПОВЕРХНОСТИ ЛИЦА ПО ОДНОМУ ИЗОБРАЖЕНИЮ

В. М. Крыгин[1], А. Н. Барановский[1]

[1] *Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского», Физико-технический институт*

Аннотация

Компьютерное зрение является активно развивающейся областью математики. Многие задачи, решение которых раньше вызывало огромные трудности, в наши дни решаются достаточно быстро в связи с развитием вычислительной математики и вычислительной техники. Одной из актуальных проблем является восстановление трёхмерной поверхности объекта по одному изображению. В данной работе проводится обзор и анализ существующих постановок и решений задачи реконструкции пространственной конфигурации поверхности лица человека по одной фотографии. Предлагаются альтернативные постановки задачи в терминах статистической теории распознавания образов. Проводится сравнение теоретической точности разных подходов, изучаются применяемые эвристические методы, а также предлагается альтернативный подход к решению задачи.

Ключевые слова: статистическое распознавание образов, компьютерное зрение, порождающая модель лица

Введение

Байесова задача распознавания была введена Теодором Андерсоном [1] в середине прошлого века. В некоторых работах используются её частные случаи без каких-либо комментариев по поводу сделанного выбора, а иногда и вовсе без упоминания байесовой задачи.

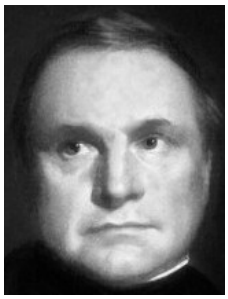

Рис. 1. Фрагмент портрета

Задача пространственной реконстукции поверхности лица по одному изображению представляет собой не только теоретический, но и практический интерес: эффективное решение можно использовать в приложениях виртуальной реальности, для биометрической идентификации пользователей какой-либо системы, в целях исторических реконструкций и так далеее.

В связи с перспективностью и важностью этой задачи было решено заново её сформулировать, а также найти слабые и сильные стороны существующих решений, чтобы предложить более оптимальный метод её решения.

1. Модель исходного изображения

На вход алгоритм получает фотографию с изображением лица. В общем случае присутствует фон и шум. Нужно описать эти факты математически, чтобы иметь однозначное понимание задачи.

Введём множество пикселей (координат) I и множество возможных цветов C. Изображение можно представить как отображение, которое каждому пикселю ставит в соответствие его цвет. Тот факт, что пиксель i изображения t принимает значение c, будем записывать как

$$t_i = c.$$

Лицо снимается при определённом освещении под определённым углом камерой с определённым объективом. Назовём эти параметры соответственно θ^L, θ^M и θ^C. Природа этих параметров в контексте данной работы нас не интересует – важно лишь то, что они заранее неизвестны. Также на предъявленном изображении помимо самого лица присутствует фон. Фон – это изображение, распределение интенсивностей пикселей которого неизвестно. Обозначим его θ^B. Нетрудно заметить, что буквой θ были обозначены параметры, которые влияют на результат синтеза изображения, но не являются частью ответа распознающей системы: нас интересует только форма поверхности лица.

Модель лица определяется набором параметров из множества $X = \mathbb{R}^n$, где n лежит в пределах от 0 до 800. Введём множество точек модели лица V. Порождающая модель лица G – линейное отображение, которое для каждого набора параметров x назначает каждой вершине v координаты в трёхмерном пространстве

$$G_v(x)_j = \lambda_0^v + \sum_{i=1}^{n} \lambda_{j,i}^v \cdot x_i, \qquad j \in \{x, y, z\}.$$

Введём отображение f, которое преобразовывает параметры x и θ в двумерное изображение поверхности лица, снятого на камеру с определённым объективом под определённым углом и освещением

$$f_i^\theta(x) = c \in C.$$

Также необходимо ввести функцию сегментации, которая равна 1 на тех пикселях, где находится лицо, и 0 на остальных

$$s : \Theta \times I \times X \to \{0, 1\}.$$

Введём аддитивный шум η, который будет накладываться на изображение модели с фоном. Пусть шум – вектор независимых случайных величин, которые имеют центрированное нормальное распределение с одинаковой неизвестной дисперсией σ_t^2.

Модель входящего изображения

$$t_i = f_i^\theta(x) \cdot s_i^\theta(x) + \theta_i^B \cdot \left(1 - s_i^\theta(x)\right) + \eta_i.$$

2. Байесова задача распознавания

Задача состоит в нахождении такого решающего правила, чтобы при данной статистической модели распознаваемого объекта минимизировалось математическое ожидание данной функции потерь. Иными словами, построенная распознающая система в среднем должна ошибаться меньше, чем любая другая.

Рис. 2. Результат реконструкции

Вероятность того, что на данной картинке t изображено лицо с параметрами x и θ, считается из тех соображений, что разность интенсивности пикселя i входящего изображения t и сгенерированного изображения t' в итоге должна быть гауссовой случайной величиной с неизвестной дисперсией σ_t^2

$$t_i - t'_i = \eta_i \sim \mathcal{N}\left(0, \sigma_t^2\right),$$

где

$$t'_i = f_i^\theta(x) \cdot s_i^\theta(x) + \theta_i^B \cdot \left(1 - s_i^\theta(x)\right). \quad (1)$$

Вероятность того, что данное изображение было сгенерировано с параметрами x и θ,

$$\mathbb{P}^\theta(t \mid x) = \prod_{i \in I} \left[\frac{1}{\sqrt{2 \cdot \pi \cdot \sigma_t^2}} \cdot \exp\left\{ \frac{\|t_i - t'_i\|^2}{2\sigma_t^2} \right\} \right]. \quad (2)$$

Априорное распределение компонент вектора x является стандартным гауссовым

$$\mathbb{P}(x) = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2 \cdot \pi}} \cdot \exp\left\{ \frac{x_i^2}{2} \right\} \right].$$

Функцию потерь следует выбирать из разумных соображений. Одной из простейших и широко применяемых является интервальная функция потерь

$$W(x, x') = \begin{cases} 1, & x \notin \delta(x'), \\ 0, & x \in \delta(x'). \end{cases} \quad (3)$$

Разумным штрафом является сумма квадратов отклонений параметров найденной модели от реальных

$$W(x, x') = \|x - x'\|^2 = \sum_{i=1}^{n} (x_i - x'_i)^2. \quad (4)$$

Также логично считать ошибку как сумму квадратов отклонений координат вершин найденной модели от реальных

$$W(x, x') = \|G(x) - G(x')\|^2 \quad (5)$$

Стратегия – отображение, которое данному изображению t ставит в соответствие параметры x

$$q(t) = x.$$

Введём понятие риска стратегии q как математическое ожидание функции потерь с использованием решающего правила q [2]

$$R(q) = \sum_{t \in T} \sum_{x \in X} \mathbb{P}(t, x) \cdot W(x, q(t)).$$

Задача распознавания состоит в отыскании стратегии, которая минимизирует риск. Поскольку входящие изображения не зависят друг от друга, можно рассматривать стратегию решения задачи отдельно для конкретного изображения [3]

$$q^*(t) = arg \min_{x' \in X} \sum_{x \in X} \mathbb{P}(t, x) \cdot W(x, x').$$

Решающим правилом для интервальной функции потерь (3) будет максимизация вероятности

$$q^*(t) = arg \max_{x \in X} \mathbb{P}(t, x). \quad (6)$$

Поскольку координаты вершин модели линейно зависят от набора параметров x, использование штрафных функций (4) и (5) приводит к одному и тому же решающему правилу

$$q^*(t) = \frac{\sum\limits_{x \in X} x \cdot \mathbb{P}(t, x)}{\mathbb{P}(t)}.$$

Данную сумму (математическое ожидание), которая в непрерывном случае переходит в интеграл, посчитать крайне трудно, так как вероятность, которая в неё входит, включает в себя сложно устроенную функцию f, аналитическое выражение которой ещё не было найдено, а потому её нельзя проинтегрировать аналитически. В связи с этим нужно использовать численные методы интегрирования, которые не дадут достаточно точного результата за приемлемое время.

Исследователи Шлезингер и Водолазкий показали, что стратегия, которая не является байесовой, является непригодной – то есть, для любой небайесовой обязательно найдётся такая байесова стратегия, которая даёт меньший риск при любых параметрах θ [4]. Байесова стратегия q при данном наборе

весов $\tau(\theta)$ минимизирует выражение

$$\sum_{\theta\in\Theta}\tau(\theta)\sum_{x\in X}\sum_{t\in T}\mathbb{P}^{\theta}(t,x)\cdot W(x,q(t))\to\min_{q}. \quad (7)$$

Функция τ ищется отдельно сообразно условиям, которые наложены на искомую стратегию. Также должно выполняться

$$\begin{cases}\sum\limits_{\theta\in\Theta}\tau(\theta)=1,\\ \tau(\theta)\geq 0,\forall\theta\in\Theta.\end{cases}$$

3. Эвристические методы

3.1. Метод максимального правдоподобия

Задача (7) является вычислительно сложной. Помимо суммы с огромным количеством слагаемых есть неизвестная функция τ, поиск которой – нетривиальная задача.

Вместо этого используется метод максимального правдоподобия. Выбирается то значение неизвестного параметра θ, при котором совместная вероятность изображения и параметров модели наибольшая

$$\mathbb{P}^{\theta}(t,x)\to\max_{\theta}.$$

Рис. 3. Другой поворот

Далее найденное значение используется так, будто это и есть реальное значение неизвестного θ.

Схожий подход состоит в том, чтобы искать параметр θ вместе с x в ходе работы определённого алгоритма. Находятся такие значения θ и x, которые являются решением задачи, а затем во внимание берётся лишь x. Из (6) получаем задачу

$$\mathbb{P}(t,x)\to\max_{x,\theta}.$$

Вероятность принимает неотрицательные значения меньше единицы, потому её максимизация эквивалентна максимизации её логарифма. Избавившись от констант и умножив на -2, получим минимизацию (из-за смены знака)

$$|I|\cdot\ln\sigma_t^2+\sum_{i\in I}\frac{\|t_i-t_i'\|^2}{\sigma_t^2}+\sum_{i=1}^{n}x_i^2\to\min_{x,\theta}. \quad (8)$$

Если считать, что дисперсия шума известна заранее, то её можно отбросить как константу

$$\sum_{i\in I}\frac{\|t_i-t_i'\|^2}{\sigma_t^2}+\sum_{i=1}^{n}x_i^2\to\min_{x,\theta}.$$

Далее для краткости вместо t_i-t_i' будем использовать y_i.

3.2. Фон

В работах, где описаны методы реконструкции пространственной конфигурации поверхности лица человека по фотографии, упускается из рассмотрения такая важная деталь, как фон. Тем не менее, исследователи прибегают к эвристикам для решения этой проблемы.

Проблема заключается в том, что возможных фонов огромное множество. Можно сгенерировать $2^{8\cdot3\cdot10}\approx10^{70}$ разных цветных изображений площадью всего лишь 10 пикселей. К тому же, фон – параметр, который на выходе нас интересует меньше всего.

Чтобы частично преодолеть возникшие препятствия, анализируются лишь те пиксели входного изображения, на которых находится лицо с данными параметрами θ и x

$$I'=\left\{i\;\middle|\;i\in I,s_i^{\theta}(x)=1\right\}.$$

Теперь в формуле (1) можно избавиться от фона

$$t_i'=f_i^{\theta}(x),i\in I'.$$

В общем случае при изменении искомых параметров изменяется мощность множества I'. Поскольку алгоритмы математической оптимизации подразумевают изменение параметров целевой функции, этот факт необходимо рассмотреть подробно.

Проблема в том, что при решении задачи (8) выгоднее всего взять такую модель, чтобы множество I' содержало как можно меньше пикселей. Если выбрать такой масштаб, что лицо сожмётся в один пиксель, или же вовсе пропадёт, сумма примет минимальное значение из возможных.

Есть два основных метода обхода этой трудности:

1) на каждой итерации случайным образом выбирать фиксированное количество пикселей [5] (стохастический градиентный спуск [6]);
2) делить сумму квадратов отклонений на количество пикселей [7].

Преимущество первого подхода состоит в его скорости. Также он ясен на интуитивном уровне, так как напоминает метод Монте-Карло. Второй метод вызывает сомнения, так как деление на переменную величину не может пройти бесследно. Грубое, но интуитивно понятное доказательство корректности таких действий состоит в том, что деление суммы квадратов отклонений на их количество даст смещённую оценку дисперсии. Если достаточно большое количество пикселей занято лицом, то оценка дисперсии на них будет близка к оценке дисперсии на всём изображении. Значит, можно посчитать её на множестве I', умножить на размер изображения, и это тем меньше повлияет на точность, чем больше $|I'|$

$$|I|\cdot\ln\sigma_t^2+\frac{|I|}{|I'|}\cdot\sum_{i\in I'}\frac{\|y_i\|^2}{\sigma_t^2}+\sum_{i=1}^{n}x_i^2\to\min_{x,\theta}. \quad (9)$$

3.3. Расположение лица

Как уже было сказано выше, из-за отсутствия статистической модели фона нельзя сказать точно, находится ли в данном сегменте изображения лицо, или же это фон. Наглядный пример – фото в комнате с чёрными обоями. Если запустить алгоритм поиска глобального минимума функции (9), то не исключено, что вместо перемещения головы на своё

место будет подобрано такое освещение, что лицо будет чёрным как стена. С точки зрения алгоритма это будет хороший минимум. С точки зрения человека такой результат никуда не годится.

Байесова стратегия (7) теоретически не подвержена такой слабости. Веса $\tau(\theta)$ будут тем меньше, чем ниже качество распознавания x при определённых θ.

В работе [5] оператор-человек грубо оценивал такие параметры как положение, поворот и освещение лица θ'. Предполагалось, что погрешность такой предобработки имеет центрированное нормальное распределение с неизвестной дисперсией σ_θ^2

$$\sum_{i\in I'\subset I}\frac{\|y_i\|^2}{\sigma_t^2}+\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{m}\frac{\|\theta^m-\theta'^m\|^2}{\sigma_{\theta^m}^2}.$$

Современный подход [7] использует опорные точки лица L', которые обнаруживаются с помощью алгоритма «выравнивание лица за одну миллисекунду» [8]

$$\omega_c\cdot\sum_{i\in I'\subset I}\frac{\|y_i\|^2}{\sigma_t^2}+\omega_r\cdot\sum_{i=1}^{n}x_i^2+\omega_l\cdot\sum_{i=1}^{m}\frac{\left\|L^\theta(x)^m-L'^m\right\|^2}{\sigma_L^2}.$$

Веса ω_c, ω_r и ω_l выбираются равными 1, $2.5\cdot10^{-5}$ и 10 соответственно по причинам, которые в оригинальной статье не описаны. Если же предположить, что решалась байесова задача распознавания с интервальной функцией потерь (6), выясняется, что эти коэффициенты соответствуют очень малым значениям дисперсий

$$\begin{cases}\sigma_t^2 &= 2.5\cdot10^{-5}\cdot|I|\,,\\ \sigma_L^2 &= 2.5\cdot10^{-6}\cdot|L|\,.\end{cases}$$

Экспериментально было проверено, что опорные точки, которые находятся исключительно в середине лица (края глаз, губ, кончик носа и так далее), иногда позволяют алгоритму сходиться на неправильном масштабе и повороте. Стандартный используемый набор точек содержит ещё и границы лица – то есть, в некотором смысле решается задача сегментации.

4. Возможные улучшения

Использование оценки области лица s' поможет

1) использовать сумму квадратов отклонений в качестве целевой функции;
2) позиционировать модель лица по силуэту;
3) учитывать перекрытие лица посторонними объектами.

Нужно, чтобы силуэт найденного лица был как можно больше похож на силуэт изображённого лица, и чтобы само лицо было как можно больше похоже на оригинальное. Задача оптимизации с ограничениями

$$\begin{cases}\sum\limits_{i:\ s'_i=1}\frac{\|t_i-t'_i\|^2}{\sigma_t^2}+\sum\limits_{i=1}^{n}x_i^2\to\min\limits_{x,\theta},\\ \sum\limits_{i\in I}\left|s_i^\theta(x)-s'_i\right|<c,\qquad c>0.\end{cases}$$

Выводы

Задача реконструкции пространственной конфигурации лица человека по фотографии не до конца сформулирована, а потому её существующие решения содержат неточности. Вычислительная сложность некоторых процедур на сегодняшний день не позволяет использовать корректные алгоритмы решения, что приводит к использованию разного рода эвристических подходов.

Данная работа содержит простейшую постановку этой задачи без учёта использования светофильтров, размытостости изображения, перекрытия лица другими объектами, теней от посторонних объектов и так далее. Даже в такой простой формулировке решение задачи сталкивается не только с техническими, но и теоретическими трудностями.

Рис. 4. Другая тень

В совокупности с экспериментальными результатами изложенная в данной работе теория дала возможность создать теоретическую основу для более точного алгоритма, который заключается в использовании сегментации для получения дополнительной информации об изображённом лице.

Перечень использованных источников

1. Андерсон Т. В., Гнеденко Б.В. Введение в многомерный статистический анализ. — Гос. изд-во физико-математической лит-ры, 1963. — С. 500.
2. Berger J.O. Statistical Decision Theory: Foundations, Concepts, and Methods. Springer Series in Statistics. — Springer New York, 1980. — P. 428.
3. Schlesinger M.I., Hlavác V. Ten Lectures on Statistical and Structural Pattern Recognition. Computational Imaging and Vision. — Springer Netherlands, 2002. — P. 522.
4. Schlesinger M. I., Volodazkiy E. V. Nearly Optimal Statistical Recognition and Learning // Proceedings of 4th International conference on Inductive Modelling. — 2013.
5. Blanz V., Vetter T. A morphable model for the synthesis of 3D faces // Proceedings of the 26th annual conference on Computer graphics and interactive techniques. — 1999. — P. 187–194.
6. Jones M. J., Poggio T. Multidimensional morphable models. — 1998. — Jan. — P. 683–688.
7. Face2Face: Real-time Face Capture and Reenactment of RGB Videos / J. Thies, M. Zollhöfer, M. Stamminger et al. // Proc. Computer Vision and Pattern Recognition (CVPR), IEEE. — 2016.
8. Kazemi Vahid, Sullivan Josephine. One Millisecond Face Alignment with an Ensemble of Regression Trees // Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. — 2014. — P. 1867–1874.