

Описание программной части робота-художника

Латыпов Владимир Витальевич

30 июля 2021 г.

Содержание

1	«Abstract»	4
2	Формулировка задачи	4
2.1	Аналог №1	4
2.2	Аналог №2	5
3	Технические аспекты	8
3.1	Основное	8
3.2	Библиотеки	8
3.2.1	OpenCV	8
3.2.2	Pythonic	8
3.2.3	lunasvg	9
3.2.4	PowerfulGA	9
4	Описание программы в общих чертах	9
4.1	Представление «решения» — набора мазков	10
4.2	Задание функции ошибки	10
4.3	Растеризация мазков	11
4.4	Проведение операций ГА	16
4.4.1	Первоначальная генерация популяции	18
4.4.2	Мутация и корректировка	18
4.4.3	Скрещивание	18
4.5	Учёт цвета при оптимизации	18
4.5.1	Знание цвета при рендеринге необходимо	18
4.5.2	Определение цвета по положению при подсчёте ФО	18
4.5.3	Фиксированный цвет в зонах	19
4.6	Разделение картины на зоны	19
4.6.1	Обоснование эффективности	19
4.6.2	Прямоугольные зоны	20
4.6.3	Зоны произвольной формы	20
4.6.4	Распределение ресурсов по зонам	20
4.7	Регуляция кривизны мазков	21
4.8	Сортировка мазков перед выпуском	23
4.8.1	Сортировка по территориальному признаку	23
4.8.2	Оптимальная расстановка цветов	24
4.9	Получение итоговых цветов: сжатие палитры	25
4.10	Соотнесение параметров с физическими величинами	25
4.11	Учёт специфики кисти	25
5	Алгоритмы оптимизации	26
5.1	Общий принцип ГА	26
5.2	Термины	26
5.3	Примерная последовательность действий ГА	26
5.4	Конкретная реализация ГА и авторские модификации	27
5.5	Операция «скрещивание»	28
5.5.1	hazing_percent: скорость сходимости	28
6	Грядущие улучшения в ГА	29

7	Наблюдения	29
7.1	Неравенство зон	29
8	Дальнейшее развитие	31
8.1	Внедрить быстрый пересчёт функции ошибки	31
8.2	Разделение мазков по слоям	32
8.3	Добавить возможность использования локальных методов оптимизации	32
8.4	Организовать систему тестирования различных алгоритмов на различных функциях	32
8.5	Контроль уровня разнообразия особей в ГА	33
8.5.1	Задание метрики	34
8.5.2	Определение требуемой динамики разнообразия	35
8.5.3	Регуляция разнообразности	35
8.6	Улучшить алгоритм поиска цветов и разделения на зоны	36
8.7	Перенести графические вычисления на видеокарту	37
8.8	Выделение границ мазками	38

1. «Abstract»

Зададимся вопросом: «Может ли робот создать произведение искусства?» На этот вопрос человечеству придётся найти ответ, и это непременно будет происходить в ближайшем будущем по мере нашего приближения к AGI (Artificial General Intelligence).

Но оставим этот вопрос будущим мыслителям и постараемся приблизиться к ответу. Так возникла идея создания робота-художника. По задумке робот должен сам придумывать что и как ему рисовать, а затем воспроизводить это кистью и красками на холсте.

Но такая обширная задача разбивается на отдельные подзадачи: само механическое устройство, его непосредственное управление и координация, с другой стороны - генерация изображения, которое необходимо нарисовать.

Эти две части связывает следующая задача: разбиение данного растрового изображения на мазки, то есть движения робота. Задача является ключевой в проекте, и именно ей занимаюсь я в данный момент. Параллельно в проекте уже реализован первый прототип механической части и алгоритм его управления. Видео тестового запуска доступно по ссылке: https://youtu.be/V8-YITMag_I.

2. Формулировка задачи

Данное описание посвящено именно задаче разбиения данного изображения на мазки. На вход подаётся растровое изображение (набор пикселей), на выходе необходимо получить набор «мазков», которые подаются на робота.

Мазки решено было представлять в виде кривых Безье второго порядка (то есть квадратичных), к которым добавлены параметры «толщина» и «цвет», сама кривая Безье задаётся тремя точками на плоскости.

Таким образом, необходимо найти такую комбинацию мазков, которая бы лучше всего соответствовала исходному изображению. Это задача нетривиальная, имеющая множество решений.

В мире уже существуют алгоритмы для решения подобных задач, однако, на мой взгляд, в них было уделено незаслуженно мало внимания вопросам оптимизации, и потому качество подбора мазков возможно было значительно улучшить. Например:

2.1. Аналог №1

A robotic system for interpreting images into painted artwork (by Carlos Aguilar & Hod Lipson)

Авторы использовали генетический алгоритм, без дополнений, которые есть у меня.

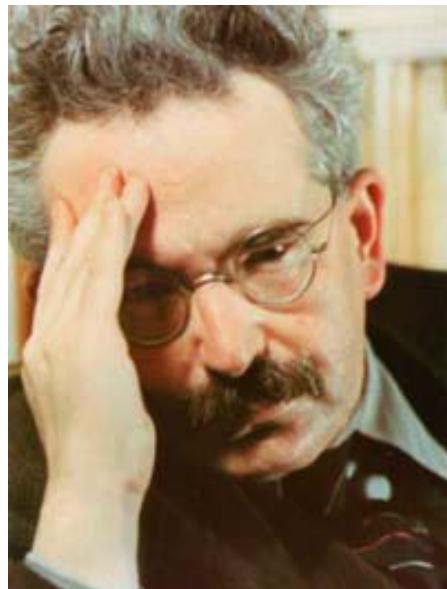


Рис. 1: Изначальное изображение, использованное авторами (1)



Рис. 2: Изображение, переведённое в мазки авторами, в электронном виде (1)

2.2. Аналог №2

Artistic Style in Robotic Painting (by Ardavan Bidgoli, Manuel Ladron De Guevara, Cinnie Hsiung, Jean Oh, Eunsu Kang)

Авторами использовалась нейросеть для оптимизации.



Рис. 3: Изначальное изображение, использованное авторами (2)



Рис. 4: Изображение, переведённое в мазки авторами, в электронном виде (2)

Здесь мы видим, что количество мазков мало, из-за чего картина лишь отдалённо похожа на исходное изображение. Но, как будет показано далее, качество оптимизации напрямую влияет на максимальное количество мазков, которое можно себе позволить, причём при увеличении количества мазков время работы очень быстро растёт, что требует существенного роста качества оптимизации.

Исходя из отчётов об их работе именно это было сдерживающим фактором на пути количества мазков (а значит, и детализации).

Пример результатов работы моего алгоритма (генерируются примерно за 5 минут):

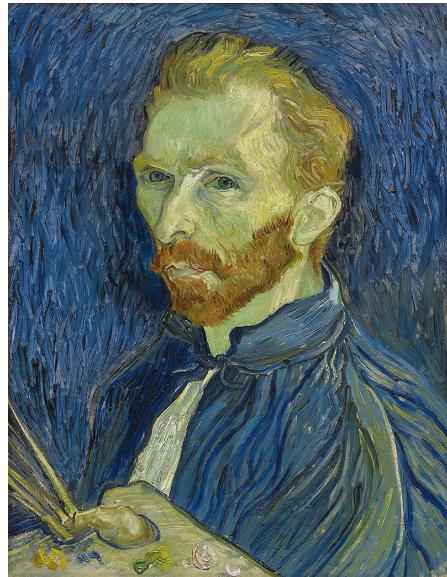


Рис. 5: Изначальное изображение (автопортрет Ван Гога)

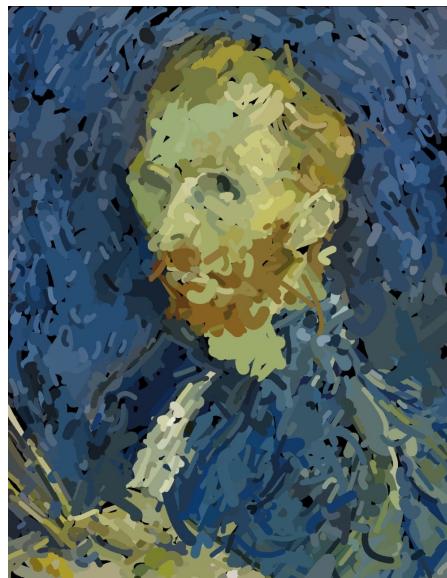


Рис. 6: Большой размер мазков

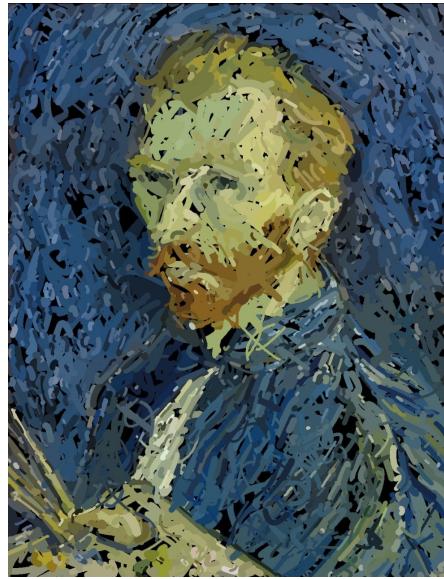


Рис. 7: Меньший размер мазков

Использовать стороннюю библиотеку для оптимизации не целесообразно, так как это существенно сужает пространство для манёвра и не позволяет углубиться в эту тему.

Поэтому для этого использовалась библиотека, написанная мной (см. подробнее в 3).

3. Технические аспекты

3.1. Основное

Программа написана на языке программирования C++ (так как требовалась максимальная скорость), сборка осуществляется с помощью CMake. Проект можно скомпилировать под Windows (компилятор MSVC) и под Linux (тестировалось на g++-10).

Код хранится в github-репозитории: <https://github.com/donRumata03/Painter>.

3.2. Библиотеки

3.2.1. OpenCV

Для работы с изображениями используется OpenCV, но не модуль машинного обучения, а лишь примитивные операции с изображениями: прочтение из популярных форматов, сохранение в них, хранение и копирование матрицы пикселей и т.д.

3.2.2. Pythonic

Для работы проекта также необходима библиотека «pythonic» (<https://github.com/donRumata03/pythonic>): она написана мной, подключается также через CMake. Она отвечает за базовые функции и структуры данных. Я использую её во всех более или менее крупных проектах на C++. В ней на данный момент есть:

- Простые вспомогательные функции для работы со строками, контейнерами, форматированного вывода
- Вызов питоновской библиотеки `matplotlib` для построения графиков
- Базовые алгоритмы наподобие бинарного поиска и дерева отрезков
- Функционал для работы со временем, в том числе — анализатор последовательных запусков процесса
- Платформонезависимая работа с кодировками и файловыми системами
- Примитивы для вычислительной геометрии
- Функции для работы со статистикой
- Многомерный шаблонный массив с количеством измерений, изменяемом в `run-time`
- Сглаживание функций и построение примерной функции распределения в пространстве с заданной размерностью по набору `sampl`-ов с помощью гауссовых ворот
- Функционал для работы с многопоточностью, в том числе — `thread pool`, умеющий снимать нагрузку с ожидающих потоков с помощью `std::condition_variable`.

3.2.3. `lunasvg`

Для работы с SVG используется библиотека `lunasvg` (<https://github.com/sammycage/lunasvg>).

P. S. У этой библиотеки отличный автор (<https://github.com/sammycage>), он изучает проекты, в которых библиотека используется, и пишет рекомендации о best practice её использования.

3.2.4. `PowerfulGA`

Функционал по методам оптимизации реализован мной (тоже на C++) и вынесен в отдельную репозиторию: <https://github.com/donRumata03/PowerfulGA> (там не только Генетический алгоритм, как можно было подумать из названия, но и симулляция отжига, градиентный спуск, метод Ньютона; планируется добавить много других алгоритмов) Более подробное описание в секции → 5

4. Описание программы в общих чертах

Задача хорошо сводится к оптимизационной задаче с большим количеством параметров. Причём на обработку одного изображения можно выделить много времени и вычислительных ресурсов — главное — результат.

Таким образом, решено было использовать эвристические алгоритмы оптимизации: «Генетический алгоритм» (ГА) и «Симулляция отжига». ГА хорошо справляется с избеганием совсем локальных оптимумов и нахождением более хороших решений, а отжиг может улучшить приближение, полученное с помощью ГА за счёт более высокой скорости сходимости (используется отжиг, а не, например, градиентный спуск, так как считать градиент в данном случае затруднительно (но его

поддержка планируется, см. 8.3)). (Исследования, которые я читал, показывают, что для большого количества стандартных задач комбинаторной оптимизации ГА показывает лучшие результаты среди алгоритмов оптимизации (ссылка есть в секции 5)) Подробное описание алгоритмов, моей их реализации, и улучшений представлено в секции 5.

4.1. Представление «решения» — набора мазков

Программа работает с последовательностями мазков, находя лучшую из них. Однако, чтобы алгоритм оптимизации работал с ними, наборы мазков должны быть представлены в виде векторов в \mathbb{R}^n .

Каждый мазок является кривой Безье второго порядка:

$$\begin{cases} x(t) = p_{1x} + (1-t)^2 \cdot (p_0 - p_1)_x + t^2 \cdot (p_2 - p_1)_x \\ y(t) = p_{1y} + (1-t)^2 \cdot (p_0 - p_1)_y + t^2 \cdot (p_2 - p_1)_y \end{cases} \quad (1)$$

И задаётся с помощью семи параметров — вещественных чисел. Формат данных в «геноме» таков:

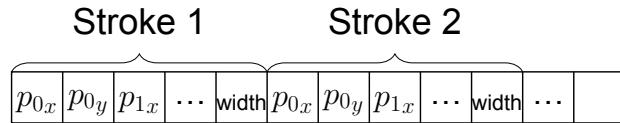


Рис. 8: Схема хранения генома

, где $p_{nx} \& p_{ny}, n \in \{0, 1, 2\}$ — координаты направляющей точки под номером n (из всего 3 у каждого мазка), а $width$ — толщина мазка.

Здесь нет параметра «цвет». Причина объяснена в разделе: 4.5.

4.2. Задание функции ошибки

Чтобы решить задачу алгоритмом оптимизации, нужна некая метрика — функция, которая будет определять степень «неподходящести» данного её решения. Именно она будет передаваться алгоритму оптимизации. В нашем случае вычисление функции ошибки (далее — ФО) включает в себя растеризацию мазков (отображение их на изображении) и вычисления, производящие сравнения полученного результата с желаемым. ФО необходимо задать таким образом, чтобы она отражала качество полученной комбинации мазков, причём в любой точке направление её уменьшения соответствовало направлению улучшения результата. За основу была выбрана MSE (Mean Square Error):

$$MSE = \frac{1}{width \cdot height} \cdot \sum_{y=0}^{y < height} \sum_{x=0}^{x < width} \sum_{c \in \{r,g,b\}} \left(\overrightarrow{\text{rendered}_{x,y,c}} - \overrightarrow{\text{original}_{x,y,c}} \right)^2 \quad (2)$$

MSE — универсальная мертика для схожести изображений, она повсеместно используется при работе с ними. Но в нашем случае, о чём свидетельствует практика, целесообразно добавить в ФО компоненту, «наказывающую» за наложение мазков друг на друга, а также за пустые (ничем не закрашенные) места.

Первое улучшение очевидно — при прочих равных лучше ситуация, при которой та же картина достигнута с меньшим использованием краски (а если не вводить эту компоненту, наложено в найденном решении может быть сразу много (> 2) мазков в одной точке). Если нет разницы, зачем переплачивать?

С теоретической точки зрения может быть непонятна надбавка за пустоты: ведь если место пустое, оно и так не даёт оптимальное MSE. Однако на практике пустоты недостаточно быстро и полно покрываются (особенно — на ускоренном режиме) без этой надбавки.

Таким образом, функция ошибки сделана так, чтобы максимально стимулировать правильно распределение мазков.

4.3. Растеризация мазков

Имея мазок, заданный в виде трёх точек на плоскости, толщины и цвета, нужно уметь его отобразить его на «холсте», то есть в виде набора пикселей. Это нужно, чтобы подсчитать функцию ошибки для заданного набора мазков, причём так, чтобы результат максимально соответствовал мазку, рисуемому роботом. В качестве достаточно точной модели описания такого мазка возьмём круглую кисть, перемещающуюся по заданной траектории. Есть много способов произвести растеризацию. Нужно выбрать тот, который будет производительным и в то же время максимально близким к реальному мазку.

Самый простой — для некоторого количества точек на кривой Безье (с достаточно маленьким шагом, примерно один пиксель) проводим вертикальную линию: вверх на width и вниз — тоже. Это даёт высокую производительность и сносно выглядит на участках, близких к горизонтальным, но результат, полученный таким способом, очень далёк от реальности на вертикальных участках:

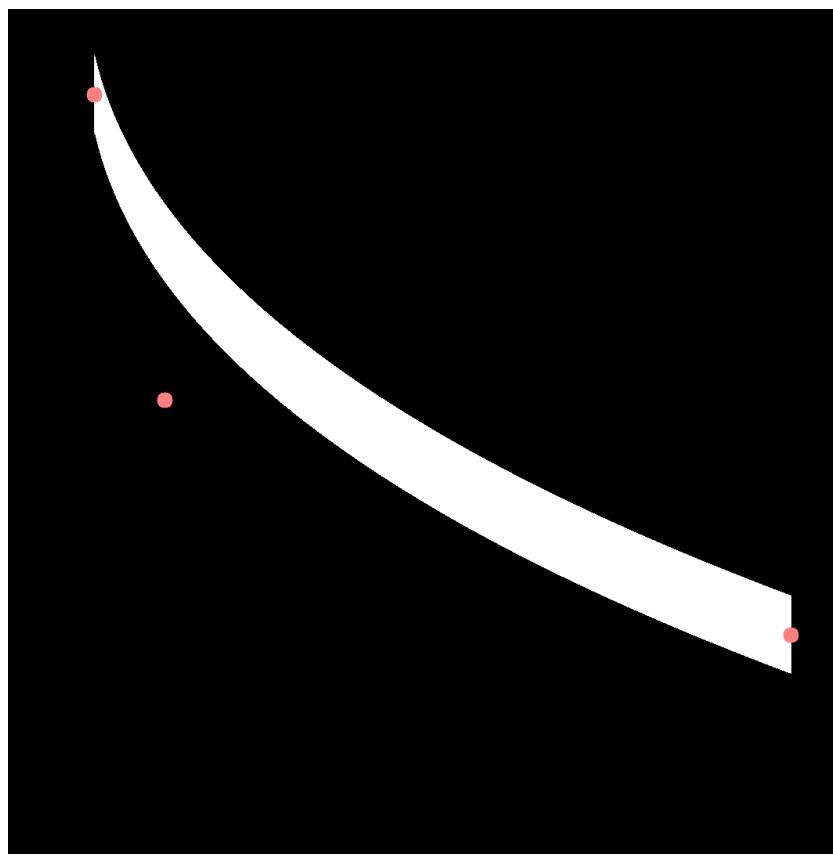


Рис. 9: (Красным обозначены точки, задающие кривую)



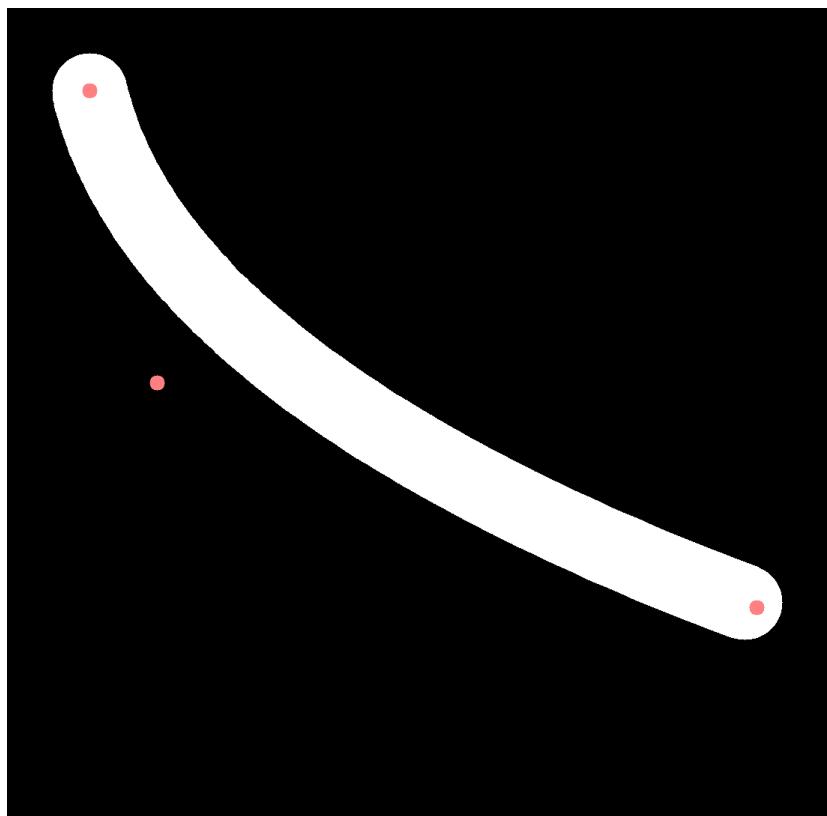
Рис. 10: Иногда такой способ добавляет свой шарм

Есть разные способы избавиться от этих недостатков, сохраняя максимальную производительность. Например:

- Совмещать горизонтальные и вертикальные полосы
- Проводить полосы перпендикулярно направлению кривой в данной точке

В каждом из них будут наблюдаться пустые места, полости, что недопустимо.

Ультимативным же способом является подражание реальной жизни: «проведение» круглой «кистью» по экрану. То есть берутся точки на кривой на небольшом расстоянии друг от друга, далее из каждой рисуется круг радиусом `width`. Однако в таком случае каждая точка, попадающая в мазок, обрабатывается много раз (для близких кругов), что значительно замедляет рендеринг. Если же увеличить шаг, этой проблемы можно частично избежать, но мазок стал бы неровным. При маленьком шаге это выглядит так:



В будущем планируется улучшить алгоритм для ускорения растеризации при почти том же качестве. Рассматриваются варианты:

- Заменить круглую кисть на также гладкую, но с более медленным закруглением с дальней от вектора кривой в данной точке стороны, поворачивая кисть соответствующим образом:

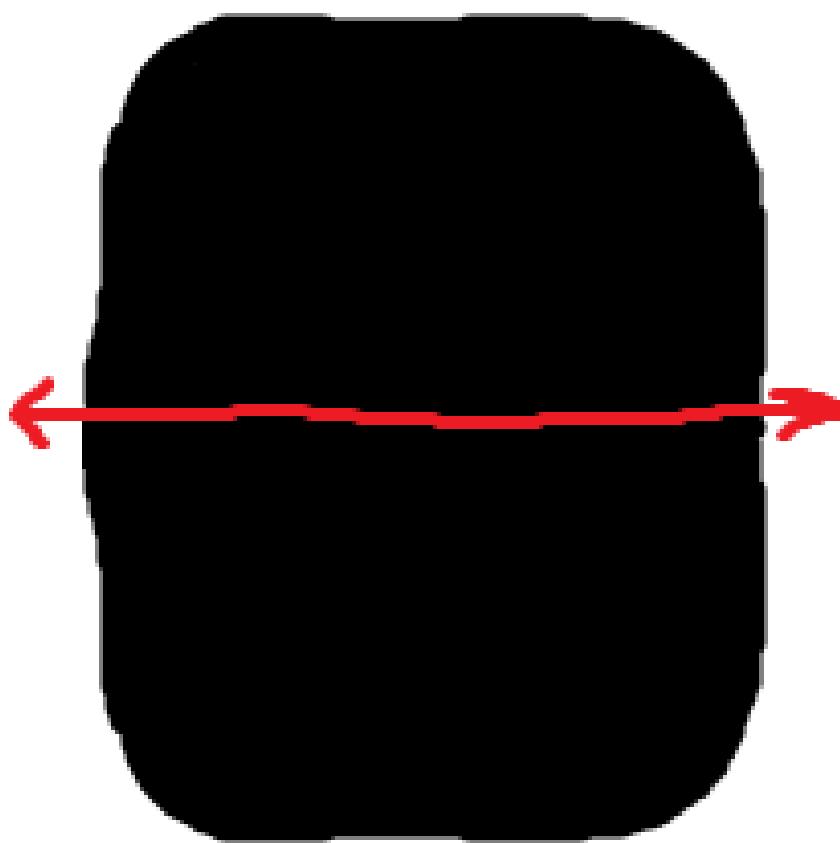


Рис. 11: (красным обозначено направление вектора кривой)

Такое изменение поможет уменьшить артефакты при увеличении шага между точками на кривой, то есть позволит сделать шаг больше, ускорив процесс.

- Автоматически разбивать мазок на «полигоны». Для этого нужно пройтись по кривой и с некоторым шагом (уже побольше, чем раньше), отметить для каждой рассматриваемой точки на прямой, содержащей её и перпендикулярной текущему направлению, точки в обе стороны от неё на расстоянии `width`. Каждая из них добавляется в соответствующий список. Потом полигоны, полученные из соседних точек на кривой и соответствующим им вынесенным точкам, заливаются нужным цветом. На концах же мазка рендерятся круги.

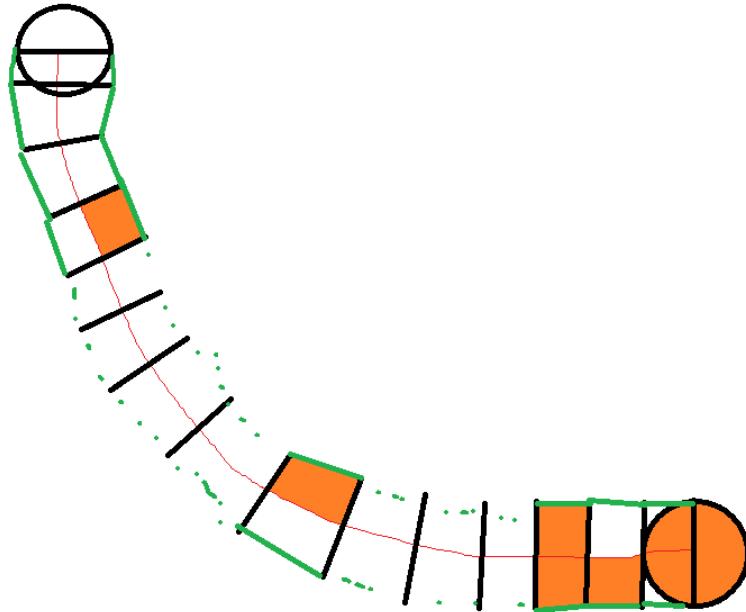


Рис. 12: Схема полигональной разбивки мазка

При использовании этого метода никакая существенная часть пикселей мазка не обрабатывается много раз, что говорит о высокой эффективности алгоритма. Поэтому я собираюсь внедрить такой метод в ближайшее время.

Что же касается совместимости с видеокартой, для круга и модифицированной кисти легко определить *bouding-box*, и несложно по координатам пикселя быстро определить, принадлежит ли он этому примитиву. А рендеринг полигонов производится аппаратно.

Подробнее про внедрение видеокарты можно почитать в разделе 8.7.

4.4. Проведение операций ГА

Задачу можно представить чисто математически — как оптимизации функции $\mathbb{R}^n \leftarrow \mathbb{R}$. Несмотря на то, что в пределе такой подход приведёт к некоторому результату,

- Это займёт очень много времени
- Результат не будет соответствовать некоторым критериям из реальной жизни

Операции, по умолчанию производящиеся в ГА, несколько изменены и адаптированы для конкретной задачи.

Важная проблема решения «в лоб» (которую нужно устраниć именно так) — то, что, если разрешить в качестве области поиска по каждой координате указать весь диапазон изображения по этой координате, типичный мазок будет растянут на всю картину. Но, во-первых, настоящие мазки — совсем не такие — они существенно меньше изображения. Во-вторых, понятно, что алгоритму будет несравненно более сложно найти нужную комбинацию мазков. Например, потому, что она, скорее всего, будет предполагать малый размер мазков, который, конечно,

возможен при такой постановке задачи, однако алгоритму потребуется огромное количество ресурсов, чтобы достичь его.

Поэтому рассмотрение координат разных точек как совершенно независимых параметров — плохая идея. Я решил составить набор ограничений на разные параметры мазков:

- Минимальные и максимальные высота и ширина bounding_box-а мазка
- Минимальные и максимальные толщина и длина мазка
- Вписываемость в изображение
- Искривлённость

Для каждого параметра необходимо уметь:

- Определять, правда ли, что заданный мазок подходит по ограничения
- «Подправлять» мазок так, чтобы значения параметров пришли в норму
- Случайно генерировать набор мазков, чтобы все соответствовали критериям

Насчёт второго пункта важно отметить: мутации точек по амплитуде — в среднем заметно меньше размера самих мазков, поэтому от методов корректировки не нужно очень высокое качество (например, какие-то гарантии сохранения формы или центра или чего-то ещё): неидеальности корректировки — тоже часть мутаций — небольших случайных изменений.

В случае с координатами — рассматривается bounding-box мазка. Когда происходит корректировка по каждой из координат, если bounding-box не помещается в изображение по этой координате, мазок перемещается по ней так, что точки на «дальней» границе «отражаются» от соответствующей границы изображения, но со случайным «коэффициентом отражения» $k \in [0; 0.5]$.

А именно — новое место для них находится на перпендикуляре к краю изображению, проходящем через изначальное положение этих точек, и удалено от границы на $k \cdot d_0$, где d_0 — расстояние от границы до изначальной точки. Иначе говоря, $new_coord = image_border_{coord} + k \times (image_border_{coord} - old_coord)$. То есть чем дальше мазок вышел за границу, чем дальше его «отбросит». Однако типичная мутация мала по сравнению с размерами изображения, поэтому не стоит ожидать, что выбившиеся мазки вдруг выползут на центр (и это хорошо).

Случайный и ненулевой вес нужен, чтобы не наблюдалось скопление мазков по краям.

Затем (для подстраховки) по каждой координате каждой точки просто делается обрезка по диапазону $\{0, max_{coord}\}$.

В случае с размерами bounding_box-а точки, если что-то не так для какой-то из координат, происходит scaling точек мазка по этой координате так, чтобы он соответствовал ближайшему разрешённому значению (то есть если размер слишком большой — верхней границе ($max_{d_{coord}}$), — ($min_{d_{coord}}$)).

Толщина просто обрезается по разрешённым границам.

Длина просто считается, и, если вдруг она слишком велика или мала (хотя это маловероятно, если подходит под bounding box, а если и так и происходит, то превышение или преуменьшение всё равно незначительное) Длина кривой определяется классическим образом, по этой формуле:

$$L = \int_a^b \sqrt{x'^2(t) + y'^2(t)} dt \quad (3)$$

Применение этого к кривым бэзье можно найти здесь: <https://members.loria.fr/samuel.hornus/quadratic-arc-length.html>

Про регуляцию кривизны можно прочитать в отдельной секции: 4.7.

Пройдёмся по операциям ГА:

4.4.1. Первоначальная генерация популяции

Как было сказано выше, здесь не просто генерируются координаты задающих мазок точек по отдельности, а сначала случайно выбираются точки в прямоугольнике с размером чуть больше, чем разрешённый размер `bounding_box-a`, происходит смещение в случайное место картинки, а в конце — используется корректирующая функция.

4.4.2. Мутация и корректировка

Здесь сначала делается небольшое случайное изменение координат, а потом — корректировка по указанным правилам.

4.4.3. Скрещивание

У новой особи просто-напросто берётся часть мазков у одного родителя, а часть — у другого.

Практика показала, что проведение этой модернизации существенно повысило качество «продукта».

4.5. Учёт цвета при оптимизации

4.5.1. Знание цвета при рендеринге необходимо

Цвет является обязательным параметром мазка, без которого непонятно, как его растеризовать. Но значит ли это, что цвет обязательно должен быть одним из параметров алгоритма оптимизации? Конечно же, нет.

4.5.2. Определение цвета по положению при подсчёте ФО

Во-первых, цвет может быть легко определён, зная положение. Самый простой способ — найти среднее арифметическое цветов пикселей. Мы автоматически получим минимальное MSE, так как для каждой из цветовых компонент соответствующая часть MSE — сумма квадратичных функций (для каждого пикселя), причём «с ветвями вверх» (так как на бесконечностях она уходит в $+\infty$), которая имеет одну точку с нулевой производной — как раз в среднем арифметическом:

$$\begin{aligned} MSE_c(value_c) &= \sum_{pixel \in pixels} \left(value_c - pixel_c \right)^2 \\ &\Rightarrow \\ MSE'_c(value_c) &= 2 \cdot \left(\|pixels\| \cdot value_c - \sum_{pixel \in pixels} pixel_c \right) \end{aligned} \tag{4}$$

, где $pixel_c$ — значение канала c пикселя $pixel$.

То есть у MSE достигается производная ноль в этой точке:

$$MSE'_c(value_c) = 0 \iff value_c = \frac{\sum_{pixel \in pixels} pixel_c}{\|pixels\|} = \overline{pixel} \quad (5)$$

Следовательно, больше нигде ноль не достигается, так как функция квадратичная. Более того, это минимум, так как функция — «ветви вверху».

4.5.3. Фиксированный цвет в зонах

Во-вторых, при делении на зоны все пиксели, относящиеся к данной зоне, имеют строго заданный цвет. Подробнее об этом читать в секции 4.6.

4.6. Разделение картины на зоны

4.6.1. Обоснование эффективности

Известно, что время оптимизации до заданного качества очень сильно увеличивается при увеличении количества параметров. Причём существенно более быстро, чем линейно. Следовательно, можно получить выгоду, разделяя эти параметры каким-то образом и оптимизируя группы по отдельности.

$$F(\|parameters\|) \gg n \times F\left(\frac{\|parameters\|}{n}\right) \quad (6)$$

Вопрос в том, в каких случаях это делать можно (то есть в каких случаях качество оптимизации заметно не уменьшится при разбиении), а в каких — нет.

Надо понимать, что нельзя независимо оптимизировать близкие мазки: только их комбинация позволит понять, хорошо ли предпринятое изменение для каждого и для всех в целом. Находить положение одного, не зная положение другого, почти бесполезно. Однако, если мазки находятся далеко друг от друга, можно безболезненно произвести деление.

Пояснение: Применительно к данной задаче супераддитивность времени выполнения алгоритма по отношению к размерности входного пространства для функций в целом можно описать так: если мутация затрагивает большое количество мазков из разных сторон изображения, «картина» для функции ошибки очень зашумлена: если ФО увеличилась, определить, следствием какой именно части мутаций было то или иное изменение ФО, можно только очень «некачественно» (то есть с малой вероятностью). А следовательно, мутации будут хаотично приниматься и наслаждаться, но это не обеспечит устойчивого движение к оптимуму через комбинирование правильных мутаций. Причём выбор маленько-го количества трансформаций за мутацию — не выход: нужно именно не такое малое количество изменений за один раз, причём таких, чтобы они были рядом, то есть сильно влияли друг на друга. А совсем малое количество изменений за мутацию — плохо, так как эти переменные сильно зависят друг от друга, оптимизация, например, сначала по одной, потом по другой и т.д. приведёт не к нахождению глобального оптимума, а к застреванию в локальном. Чтобы выйти из него, нужно изменение сразу большого количества переменных. Конечно, ГА — это не Hill climbing algorithm, он не «отрезает» сразу любое изменение, не приведшее к результату, но вышеупомянутые механизмы всё равно привозят к уменьшению производительности при слишком маленьком количестве изменений за одну мутацию.

Поэтому (так как важно немалое количество изменений, сконцентрированных в одном месте, чтобы избежать попадания в локальный оптимум, — с одной стороны и отсутствие сбивания рыночных целевофункциональных сигналов другими частями картины), важно как-либо разделять изображения на зоны: функция ошибки сепарельна, но только до какого-то размера зоны, при малом количестве мазков, близких друг другу, это перестаёт работать.

4.6.2. Прямоугольные зоны

Проще всего в реализации — делить изображение на прямоугольные зоны. Каждая рассматривается как отдельное изображение, для него находятся мазки, которые потом со сдвигом добавляются «в общий котёл».

Встаёт очевидный вопрос: как избежать заметности границ этих зон на финальном изображении?

Первое средство — расчерчивать эти границы наложенными друг на друга. То, какую часть перекрытие должно составлять от всей зоны, можно настроить эмпирически, запуская программу с разными значениями этого параметра и сравнивая плотность в местах стыка и на остальной картине. Подобранное значение почти универсально, то есть может быть применено и для обработки других картин.

Второе средство — после получения результатов для зон «разблокировать положение» мазков, лежащих рядом со стыками, запустив оптимизацию для них уже на основе данных целой картины. Конечно, и здесь имеет смысл отдельно оптимизировать «крести», образующиеся при стыке четырёх зон. Также можно в качестве ограничений использовать не весь box картины, а зоны этих крестов. Более того, если окажется, что плотности в местах стыка будут не хватать, можно добавить некоторое количество новых мазков (т.н. клеящего вещества).

Последние улучшения не были внедрены, так как ↓↓

4.6.3. Зоны произвольной формы

Понятно, что лучшее качество даёт деление на зоны, связанное со структурой картинки. То есть цвет должен не сильно меняться в пределах зоны. Вопрос в том, как изображение на такие зоны поделить.

Такие программы, как Adobe Illustrator, умеют эту выполнять эту операцию (правда, с некоторыми проблемами, о них — позже: 8.6).

Adobe Illustrator генерирует .svg файл, в котором зоны представлены в виде части плоскости, ограниченной кривой Безье с большим количеством точек. В программе он парсится, затем для каждой зоны находится bounding-box, его сдвиг, и для такой прямоугольной картинки с зоной находятся мазки. Затем все мазки (с учётом сдвига) складываются в единую кучу.

4.6.4. Распределение ресурсов по зонам

Специфика задачи состоит в том, что нет чёткого момента, до которого стоит производить оптимизационные итерации: чем их будет больше, тем выше будет качество итогового продукта.

Поэтому встаёт вопрос: как именно распределить вычислительный ресурс между зонами? Такой же вопрос и с мазками. В условиях большого разброса по размеру зоны необходимо уделить этой проблеме внимание.

Самое простое решение — чем больше площадь зоны, тем больше мазков заданного размера нужно, чтобы её покрыть (зависимость линейная). А чем больше мазков, тем сложнее алгоритму оптимизации, то есть тем больше ему нужно дать вычислительного ресурса (по сути — количества вычисления целевой функции). Здесь зависимость в реальности нелинейная, скорее всего — экспоненциальная, но с небольшим основанием. Однако распределять итерации можно и линейно, чтобы не произошла ситуация, при которой на маленькие зоны не выделены итерации вообще, а не средние — почти совсем. Далее при этом подходе количества итераций и мазков, отведённые для всей картины, распределяются по зонам пропорционально их площади.

Однако требуемое количество мазков зависит не только от площади, но и от «сложности» зоны. Можно сказать, что, если у зоны больший периметр при той же площади, то она более сложная. В качестве образца «простоты» был, конечно же, взят круг — он обладает минимальным периметром при заданной площади.

После экспериментов с GeoGebra (программа для работы математическими примитивами) я составил такую систему:

Цель состоит в том, чтобы до какого-то момента увеличение периметра сильно влияло на оцениваемую сложность, а потом — нет, чтобы влияние стремилось к асимптоте.

Задаётся значение `max_perimeter_contribution` — оно определяет максимальное количество раз, в которое большой периметр может увеличить значение сложности по сравнению с «голой» площадью.

Рассчитывается:

$$\text{relative_excess} = \frac{\text{stroke_perimeter} - \text{circle_length}}{\text{circle_length}} \quad (7)$$

$$\begin{aligned} \text{perimeter_contribution} &= \text{max_perimeter_contribution} \\ &\times \left(1 - e^{\left(\frac{\text{relative_excess}}{\ln(1 - \text{characteristic_y})} \right)} \right) \end{aligned} \quad (8)$$

, где с помощью параметров `characteristic_y` и `x_of_characteristic_y` можно задать, точку на кривой, по которой определяются её (кривой) параметры. Таким образом можно регулировать, насколько периметр будет влиять на оценку сложности зоны.

И, наконец:

$$\text{complexity} = \text{zone_area} \times \text{perimeter_contribution} \quad (9)$$

Таким образом, такой метод позволяет более комплексно оценить, насколько сложна зона, и поэтому распределение пропорционально этому параметру даёт лучшие результаты.

4.7. Регуляция кривизны мазков

Если не регулировать кривизну мазков отдельно, можно заметить, что они зачастую становятся похожи не на мазки, а на георгиевскую ленту:



Изначально я собирался определять «кривость» мазка через рассмотрение точек, в которых предел отношения поворота направляющего вектора к длине пройденного шага превышает определённый порог. Я планировал изучить график зависимости этой величины от параметра t для разных мазков (тех, которые кажутся мне слишком и не слишком кривыми) и составить качественную метрику.

Однако важно не только замерять значение этого параметра, но и уметь эффективно его уменьшать, не изменяя мазок очень сильно (то есть оставив его «индивидуальные особенности»).

Тут я понял, что кривость мазка можно хорошо определять по отношению удалённости срединной из трёх задающих точек от отрезка, проведённого между начальной и конечной точками, к длине этого отрезка. Здесь важно, что расстояние от точки до отрезка — длина перпендикуляра на прямую только в том случае, если перпендикуляр попадает на отрезок, а иначе — длина до ближайшего конца отрезка.

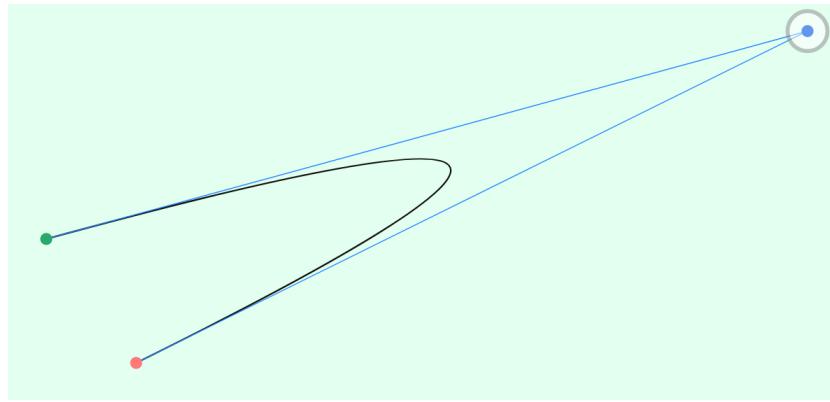


Рис. 13: «Мазок» с высокой степенью искривлённости

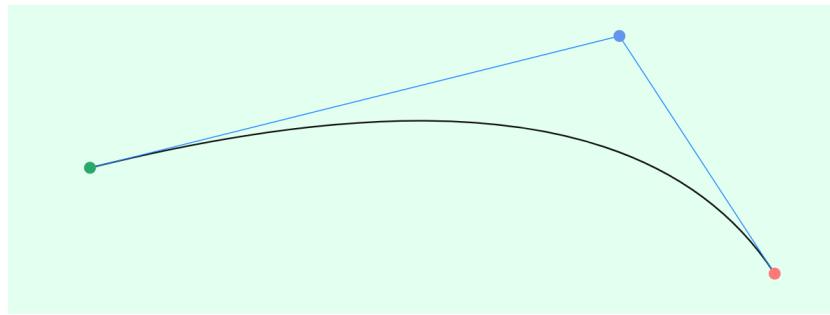


Рис. 14: «Мазок» с низкой степенью искривлённости

Отсюда вытекает и очевидный способ уменьшить кривизну: переместить срединную точку так, чтобы она была достаточно блика к отрезку. Нужно приблизить её в нужное количество раз по направлению либо к основанию перпендикуляра, либо к концу отрезка в зависимости от того, куда попадает перпендикуляр.

4.8. Сортировка мазков перед выпусктом

В случае растеризации набора мазков на компьютере время выполнения операции почти совсем не зависит от порядка следования мазков (если не считать незаметное влияние кэширования в процессоре), так как закрашивание пикселя происходит через *random access*. Однако в реальности это далеко не так: правильный порядок мазков (как по координатам, так и по цветам) может сильно уменьшить время отрисовки, которое весьма велико, и поэтому нельзя забывать про задачу правильного расположения мазков.

4.8.1. Сортировка по территориальному признаку

В первом приближении количество цветов очень небольшое (скорее всего, не больше десяти), а, чтобы сменить цвет, нужно вести кисть к банке с водой, смывать краску предыдущего цвета, брать новый, а потом опять нести кисть через весь стол. Поэтому разумно сначала разделить мазки на группы по цветам, а в каждой из них располагать их так, чтобы минимизировать пройденное кистью расстояние при последовательной отрисовке мазков этого цвета.

По сути это модификация задачи Коммивояжёра, где в роли городов выступают мазки. Разница в том, что мазки гораздо более некорректно считать матери-

альными точками: их размер — порядка расстояний между ними. То есть «вход» в мазок находится с одной стороны, а «выход» — с другой, причём для каждого мазка появляется два варианта его проведения — с одной стороны и с другой.

Конечно, можно оформить нахождение лучшей последовательности как оптимизационную задачу, где параметрами будут:

- Некая перестановка идентификаторов всех мазков
- Вектор бинарных величин для каждого мазка (с какой стороны начинать его рисование)

Но это будет неоправданно ресурсозатратно, так как от решения этой задачи зависит только скорость рисования, но не качество конечного продукта.

В качестве простого и быстрого алгоритма напрашивается обход «змейкой»: сортировать мазки сначала по одной координате, затем - по другой (в качестве якорной точки разумно использовать точку $\vec{r} = bezierCurve(t = 0.5)$ (смотреть по формуле 1)). Однако на практике первая координата примерно никогда не совпадает у нескольких мазков, поэтому сортировка по сути происходит только по ней, что приводит к постоянному метанию с одной стороны холста на другую.

Потому разумнее разделить изображение на прямоугольные зоны такого размера, чтобы количество зон было порядка количества мазков в зоне, тогда делений по одной координате: $\approx \sqrt[4]{\|strokes\|}$. Далее можно произвольно выбрать порядок мазков внутри каждой зоны: будет гарантироваться, что между каждыми мазками будет пройдено расстояние

$$\leq \sqrt{\left(\frac{image_{height}}{n_{separators}}\right)^2 + \left(\frac{image_{width}}{n_{separators}}\right)^2} \quad (10)$$

Как вариант — запустить алгоритм оптимизации внутри каждой зоны: здесь, как проверено на похожих данных в репозитории PowerfulGA возможно почти моментально получить хороший, скорее всего — идеальный — результат.

Сейчас используется разбиение на зоны с дальнейшей сортировкой по одной из координат.

Такой способ даёт удовлетворительное качество, однако рассматривается использование алгоритма Кристофидеса — возможно, он даёт более хорошее качество (гарантирует длину пути не более, чем в 1.5 раз больше минимального).

4.8.2. Оптимальная расстановка цветов

Когда внутри каждого цвета мазки расставлены, встаёт ещё один вопрос: как должны быть упорядочены сами цвета. По разным причинам предпочтительно делать это в «хронологическом», «плавном» порядке — чтобы близкие цвета сильно не отличались и мы постепенно переходили от одной крайности к другой. Это и уменьшает артефакты при использовании нового цвета после старого, и (в перспективе, когда будут масляные краски) необходимо для смешения. Более того, это полезно для получения итоговых цветов (см. 4.9)

Здесь полезно представление цвета как вектора в трёхмерном пространстве (компоненты вектора нём — это компоненты R, G и B).

Далее необходимо формализовать задачу. Есть несколько вариантов выбора целевого функционала:

- Непосредственно, длина ломанной

- Сумма квадратов длин отрезков ломанной. Этот вариант предпочтительнее, так как дополнительно наказывает за большие промежутки между цветами (то есть стимулирует равномерность в последовательности).

В первом случае мы снова получаем задачу Коммивояжёра, а во втором — просто некую оптимизационную задачу. Причём отжигом её решать уже целесообразно, так как количество цветов обычно не больше 1000.

4.9. Получение итоговых цветов: сжатие палитры

Чтобы реальный робот быстро рисовал изображения из реальной жизни (а не тестовые), количество используемых цветов должно быть невелико (скорее всего, не больше десяти). Однако сейчас генерируется большое количество цветов, и у мазков редко каких зон цвета совпадают. Нужно без потери качества «сжать» палитру — каждому изначальному цвету сопоставить один из немногих новых.

Ориентировочное количество цветов (целевой размер палитры) задаётся пользователем исходя из физических ограничений робота (на самом деле, в результате количество цветов будет в точности равно этому числу).

Имея «плавную» последовательность цветов (из 4.8.2), можно использовать её для сжатия палитры: алгоритм очевиден.

Измерим полученную длину пути от первой до последней точки (в трёхмерном цветовом пространстве) или сумму квадратов расстояний между последовательными точками (в зависимости от того, какая метрика выбрана). Назовём полученное значение L . Затем будем двигаться по последовательности цветов, поддерживая счётчик сумм длин или их квадратов, и каждый раз проверять, не достигли ли мы значения $\frac{L}{n}$, где n — количество цветов. При достижении переходить к добавлению в новую группу. Так мы разделили цвета на группы близких. Осталось в каждой выбрать цвет, наилучшим образом представляющий изначальные цвета, собравшиеся в этой группе. Как было показано в 4.5.2, для этого достаточно составить цвет из средних арифметических по каждой компоненте.

4.10. Соотнесение параметров с физическими величинами

Были проведены измерения толщин кисти, калибровка и т.д. Размеры задаются через среднее геометрическое и размер диапазона (в «разах»):

$$min_p = \frac{g_{avr}}{\sqrt{range_size}} \quad (11)$$

4.11. Учёт специфики кисти

В ходе запусков выяснилось, что робот не может рисовать короткие, но толстые мазки. Поэтому было решено изменить параметры, находящиеся в геноме. Теперь будет храниться не $width$, а $fatness$ — отсылка к тому, что для человека с бо́льшим ростом выше максимальный порог приемлемой толщины тела.

Сделано это для того, чтобы ни при каких значениях генома, входящих в указанный диапазон, не было тонких и коротких мазков.

Теперь достаточно задать трёхмерную функцию, которая по $fatness$ и рассчитанной длине мазка находит $width$. Например, при очень большой длине верхняя грань ширины для разных значений $fatness$ — это толщина кисти при максимальном нажатии.

Такой подход — принципиальная альтернатива методу с «корректировкой». В случаях, когда можно, чтобы на всём допустимом множестве значений, указанном алгоритму оптимизации, генерировалась потенциально приемлемая картинка.

Это будет реализовано в ближайшее время.

5. Алгоритмы оптимизации

5.1. Общий принцип ГА

Идея работы генетического алгоритма заимствована у природы: также, как в ходе эволюции, происходит появление оптимального организма для заданных условий — в ходе работы алгоритма ищется набор параметров, при котором фитнес-функция максимальна.

В природе тот, кто лучше приспособлен к окружающей среде, в большей степени получает доступ к размножению, в результате чего новые особи получают признаки от лучших родителей. Из-за того, что геном наследуется от обоих родителей, получившиеся комбинации могут по-новому комбинировать в себе черты, что даёт большое преимущество перед бесполым размножением (аналогом которого являются такие методы оптимизации, как градиентный спуск или отжиг). Нужно пояснить, что половое размножение не обязательно предполагает наличие нескольких полов. Особи в ГА — аналоги гермафродитов из живой природы: в отличие от людей, каждая особь может скрещиваться с каждой.

Это позволяет именно лучшим чертам, по каким-либо причинам появившимся у особей, переходить в следующее поколение.

За их появление отвечают мутации — небольшие случайные изменения в геноме.

Из принципа работы можно понять, что алгоритм эвристический: сложно доказать его сходимость или что-либо гарантировать с вероятностью 100%. Зато исследования (talgal.org/news/wp-content/uploads/2018/08/112.pdf) показывают, что именно этот алгоритм даёт лучшие результаты для самых сложных функций. В разделе 5.5.1, какие меры предпринимаются, чтобы не дать алгоритму попасть в локальный минимум, не добравшись до глобального.

5.2. Термины

Набор параметров представляется в виде «генома» — некой структуры данных, содержащей информацию об этом наборе. Особь — в контексте алгоритма будет использоваться в качестве синонима к геному.

Геном состоит из генов — каждый из них содержит информацию о каком-либо признаком (в случае природы) или параметре (в случае ГА).

В каждый момент времени алгоритм работает с популяцией — набором геномов. Это аналог популяции в природе.

Мутация — как и в реальной жизни — небольшое случайное изменение генома без строго определённого направления.

5.3. Примерная последовательность действий ГА

В общих чертах работа ГА выглядит так:

Инициализация: Сгенерировать случайную популяцию, каждый ген каждого генома — в заданных пределах.

Затем — повторять, пока не закончится заданное количество итераций или не будет достигнуто требуемое значение фитнесс-функции:

1. Посчитать фитнесс-функции для каждой из особей. Для большинства задач этот шаг занимает бо́льшую часть времени исполнения, поэтому нужно оптимизировать именно его, в частности — распараллелить, запуская независимые вычисления на нескольких потоках.
2. Каким-либо образом отобрать особи на скрещивание
3. Произвести скрещивание, получив «отпрысков» — часть нового поколения
4. Сформировать новое поколение, используя, возможно, в разных пропорциях, различные источники геномов, а именно:
 - Отпрысков, полученных на предыдущем шаге в результате скрещивания
 - Лучшие особи из прошлой популяции
 - Случайные особи — чтобы не дать алгоритму сойтись раньше времени, попав в локальный минимум
 - Возможно, результаты скрещивания одновременно более двух особей.
5. Произвести мутации в некоторых особях этого поколения (лучшие из мутаций внедряются в популяцию на следующих итерациях).
6. Опционально — «обрезать» (то есть насильственно подправить) те мутации, которые привели к выходу каких-либо параметров или их комбинаций за пределы допустимого.
7. Если алгоритм подходит к концу, добавить лучший геном из предыдущего поколения (чтобы он не подвергся мутации)

5.4. Конкретная реализация ГА и авторские модификации

Учитывая тот факт, что в большинстве задач, решаемых мною, бо́льшая часть вычислительного времени ($\gg 95\%$) используется для подсчёта функции ошибки, а не для манипуляций с геномами (это подтверждается результатами профайлинга), задача состоит в том, чтобы минимизировать количество подсчётов функции ошибки, пусть и ценой более долгой работы с геномами.

Первое изменение — отказ от дискретного кодирования геномов.

Для алгоритма по каждой переменной задан её диапазон.

Традиционный подход — разделить диапазон на 2^N частей и кодировать номер части в геноме как битовую последовательность из N бит. Для подсчёта функции ошибки этот номер перекодируется назад в соответствующую точку непрерывной величины. Мутацией в данном случае является изменение случайного количества каких-то битов этого номера.

А скрещивание обычно происходит путём случайного выбора значений каждого из двоичных разрядов в родительских геномах.

Предварительное тестирование показало бо́льшую эффективность хранения самого числа (без кодирования) по сравнению с традиционным подходом (хранение битовой последовательности), однако планируется провести более тщательное тестирование (см. 8.4)

5.5. Операция «скрещивание»

Выбор особей для скрещивания происходит пропорционально значению фитнес-функции в некоторой (небольшой) степени, что увеличивает разнообразие, так как в случае небольшой степени «оригинальные» геномы, но с неидеальной функцией ошибки могут попасть в «отцы».

Для реализации используется «рулетка» с кумулятивными вероятностями и бинарный поиск по ней при генерации новой особи по случайному числу от 0 до 1.

Само размножение (по умолчанию) осуществляется так: имея два родителя, каждый ген с некоторой вероятностью либо берётся от отца, либо от матери, либо (так как это непрерывный вещественный параметр) генерируется по такой плотности вероятности (значения параметра у двух родителей — два крайних максимума на этой кривой):

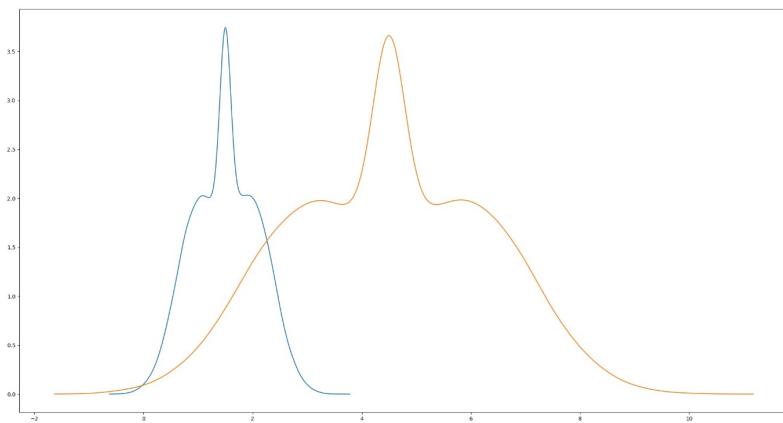


Рис. 15: Синяя кривая — для близких геномов, оранжевая — для более далёких

Кривая формируется через сумму трёх нормальных распределений.

Она было подобрано таким образом, чтобы наилучшим образом раскрывались потенциальные возможности генетического материала: максимумы, например, наблюдаются в изначальных точках и в центре между ними — это именно то, что может нам «дать» такой генетический материал.

Факт в том, что при использовании обычного бинарного кодирования просто-напросто дискредитируется изначальная природа данных, что не позволяет достичь хороших результатов.

5.5.1. `hazing_percent`: скорость сходимости

При использовании алгоритма оптимизации должен поддерживаться баланс между скоростью сходимости и избежанием застревания в локальном оптимуме.

В случае моей модификации ГА этот баланс регулируется коэффициентом `hazing_percent`.

При формировании новой популяции, как было сказано выше, используются разные категории геномов:

- Некоторое количество отпрыски, полученные на предыдущем шаге в результате скрещивания (Подробнее про это — в разделе 5.5): назовём их *children*.
- Иногда — лучшая особь из предыдущей популяции (одна) — *best_genome*.

- Некий набор «элитарных» геномов (*elite*) — отбор на эти места осуществляется пропорционально фитнесс-функции также в некой степени, но уже в большей, что делается её действительно элитной.
- Гиперэлита (*hyper_elite*) — то же самое, но меньше мест и больше степень.
- ... (Остальные)

Изменяя распределения мест по категориям, можно регулировать скорость сходимости. На это распределение и влияет параметр *hazing* («дедовщина»): высокое его значение приводит к доминированию уже сформировавшихся геномов и скорейшей их «дошлифовке», однако не даёт новым, «подающим надежды» развиться и закрепиться.

Кроме того, в каждой эпохе (итерации) определяется распределение, учитывая процент выполнения алгоритма — к концу запускается режим максимальной дедовщины: все, кто смог подать надежду, уже закрепились, осталось как раз их «дошлифовать».

(Нужно исследовать всю область поиска, поэтому нужно не давать сразу огромный бонус при размножении и переходе в другое поколение за некоторое преимущество.)

Только так получится обеспечить развитие нескольких «очагов», внутри которых и будет происходить «шлифование» «идей искать в этой области».

В будущем планируется сделать распределение мест по степеням отбора непрерывным, исследовать его, найти оптимальное (возможно — тоже с помощью метаалгоритма оптимизации).

6. Грядущие улучшения в ГА

?? О них написано в 8.

7. Наблюдения

7.1. Неравенство зон

Когда я заметил, что зоны, на которые Adobe Illustrator делит изображение, могут очень сильно отличаться в размере (отношение площадей может достигать 1000-и), мне захотелось измерить это неравенство численно, чтобы при разработке нового алгоритма измерять его качество в том числе по этому параметру. (понятно, что наличие зон слишком малого или слишком большого размера плохо сказывается на работоспособности алгоритма; то же самое можно сказать и про узкие и длинные зоны, особенно — если их диаметральная ось не близка по направлению к координатной сетке, потому что в таком случае площадь *bounding-box'a* существенно превышает площадь самого мазка)

Существует большое количество метрик, я решил выбрать основные из них:

- Индекс Джини (вместе с кумулятивным графиком распределения дохода (также известен как кривая Лоренца))
- Процент «дохода» 1% самых богатых от общего «дохода» (В случае зон вместо дохода используется занимаемая площадь).
- Процент самых богатых, имеющих в сумме 50% от общего дохода.

Индекс Джини рассчитывается как отношение площади между кривой Лоренца и «линией равенства» к площади под линией равенства. Иными словами, $G = \frac{A}{A+B}$ на этой схеме:

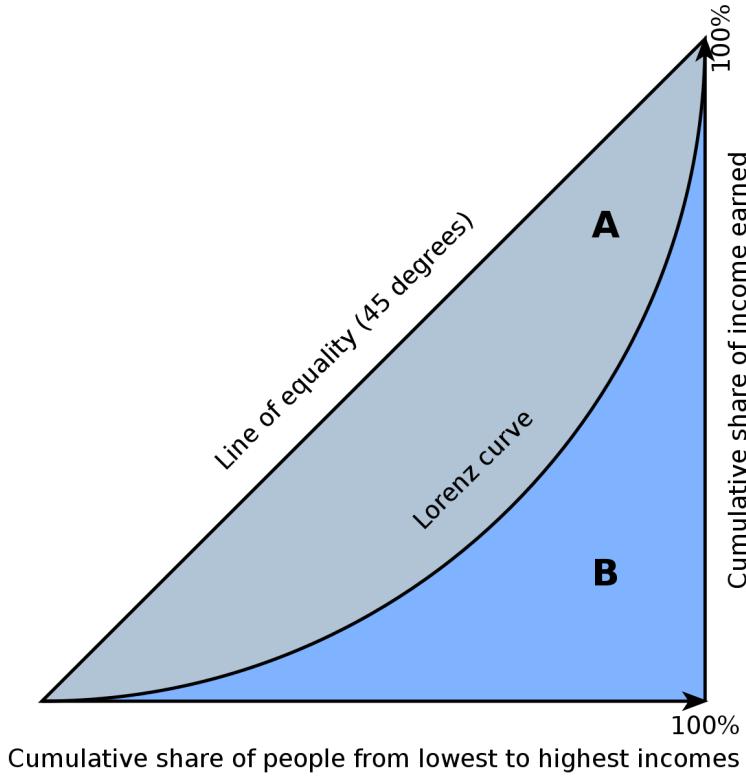


Рис. 16: Типичная кривая Лоренца

Альтернативный способ посчитать коэффициент, использующийся в реализации:

$$G = \frac{\sum_{i=1}^n \sum_{j=i+1}^n |y_i - y_j|}{n \cdot \sum_{i=1}^n y_i} \quad (12)$$

Чем индекс выше, тем большее неравенство наблюдается в стране. Более того, использование именно этой метрики позволяет комплексно оценить неравенство между анализируемыми объектами — в отличие от рассмотрения процентов дохода заданного квантиля.

Результаты оказались впечатляющими:

- $Gini_index \approx 76\%$
- 1% крупнейших зон покрывают ≈ 18% изображения
- 6.25% зон покрывают половину изображения

Так выглядит кумулятивный график распределения площади:

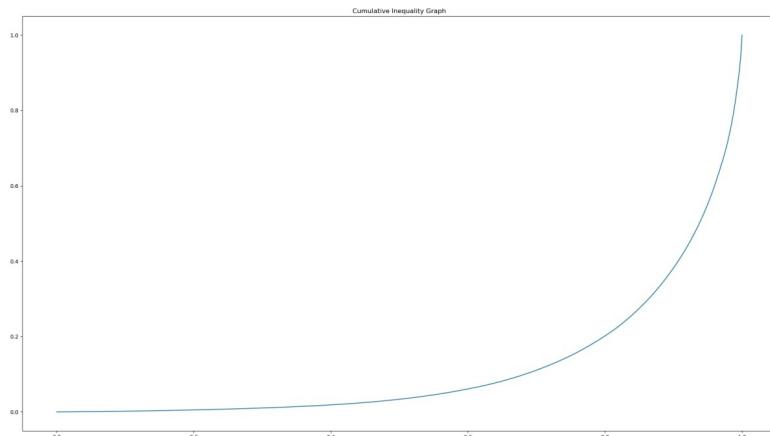


Рис. 17: Кривая лоренца для зон

Нетрудно заметить, что ни в одной стране мира нет такого неравенства, как среди зон:

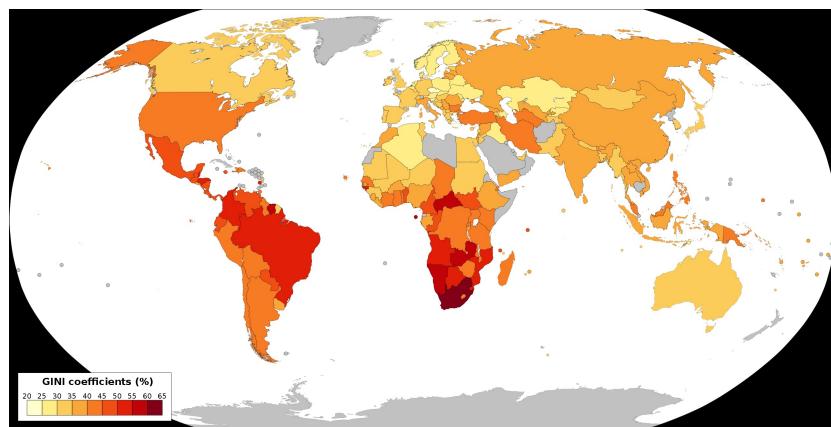


Рис. 18: Индекс Джини по странам мира

Даже в ЮАР индекс Джини составляет 57.8%.

8. Дальнейшее развитие

Несмотря на то, что программа уже работоспособна, есть ещё много идея и планов по её усовершенствованию:

8.1. Внедрить быстрый пересчёт функции ошибки

Это улучшение давно напрашивается, но оно несколько теряет в эффективности из-за того, что в одной мутации в среднем изменяется не так мало мазков (однако это количество убывает со временем). В настоящий момент ведётся работа над внедрением.

8.2. Разделение мазков по слоям

Нетрудно заметить, что при рисовании картин художники сначала проходятся по холсту черновыми мазками большого размера, а затем — прорабатывают детали. Таких уровней детализации зачастую бывает немало.

Пример того, как художник (<https://www.youtube.com/watch?v=VaXHta12alU>) рисует картину по слоям:



Рис. 19: Фон | Рельеф фона | Детализация заднего плана | Основные объекты

Поэтому стоит попробовать сначала заполнять картинку толстыми, грубыми мазками (то есть просто с большей шириной, а в реальной жизни это будет отражаться в большем размере кисти и в более сильном нажатии).

8.3. Добавить возможность использования локальных методов оптимизации

Такие методы, как градиентный спуск и метод Ньютона позволяют достичь гораздо большей скорости сходимости (в случае метода Ньютона — сходимость квадратичная (http://w.ict.nsc.ru/books/textbooks/akhmerov/mo_unicode/4.html)), но требуют умения посчитать градиент функции ошибки в любой точке, а также вектор вторых производных по каждому из аргументов.

Сами алгоритмы реализованы и находятся в этой папке: https://github.com/donRumata03/PowerfulGA/blob/master/other_optimization/. Предусмотрена опция подсчёта первой и вторых производных через подстановку близких значений параметров:

$$f'(x_0) \approx \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} \quad (13)$$

Однако в случае с мазками при маленьких изменениях параметров функция ошибки остаётся неизменной, так как это приводит к такому же набору закрашенных пикселей. Соответственно, нужно либо радикально увеличивать разрешение изображения, либо использовать аналитические методы. То есть нужно математически посчитать изменение функции ошибки при бесконечно малом изменении из параметров функции.

8.4. Организовать систему тестирования различных алгоритмов на различных функциях

Звучит как нечто весьма простое, но реальность сложнее, чем кажется. Направляющийся вариант — дать каждому алгоритму заданное количество вычислений функции ошибки и сравнить, какой результат они получат.

Однако функция ошибки нелинейная, поэтому сложно будет понять, насколько сильному различию в качестве алгоритма соответствует полученная численная разница в результатах.

Целесообразно сравнивать количество итераций, требующееся алгоритмом для получения заданного результата. Но и тут не всё так просто: нельзя просто запустить алгоритмы на неограниченное количество итераций и ждать достижения нужного значения функции, так как во многих из них (как минимум — в моей модификации ГА) то, как будет проведена каждая отдельная итерация, сильно зависит от процента выполнения на момент её прохождения: происходит планирование, использующее информацию о максимальном количестве итераций.

Поэтому нет никакого другого выхода, кроме того, чтобы запускать этот алгоритм с разным количеством итераций и смотреть, когда он в среднем будет доходить до заданного порога. Это необходимо автоматизировать.

В идеальном случае для поиска порога можно было бы использовать бинарный поиск, но в реальности (с поправкой на шум) имеет смысл использовать эвристическую модификацию бинарного поиска, а именно — на каждом шагу делить отрезок поиска на n частей, находить ту, в которой лучшее значение целевой функции, а затем переходить к рассмотрению отрезка с центром (по возможности) в этой точке и с шириной t шагов (по возможности — значит, если лучшая точка — близко к краю — немного сдвинуть новый отрезок, не изменяя его длину). Для полной оценки планируется построить график достаточного количества итераций от требуемого значения функции в интересующей нас зоне.

Это нужно проделать для нескольких сложных функций (например, из списка в Википедии https://en.wikipedia.org/wiki/Test_functions_for_optimization) и на основе этого сделать вывод об общем качестве работы.

Умение хорошо оценивать работу алгоритма «в полной комплектации» даёт возможность оптимизировать гиперпараметры. То есть мы запускаем метаалгоритм, параметры которого — это гиперпараметры основного алгоритма, а функция ошибки — качество его работы. В случае с ГА, например, гиперпараметрами могут быть: `hazing_percent`, `elite` или `hyperelite_fit_pow`, `epoch_pow`, `mutation_percent_sigma`, `target_gene_mutation_number` и т.д.

8.5. Контроль уровня разнообразия особей в ГА

Известно (5.5.1), что в ГА важен баланс между скоростью сходимости и разнообразием в популяции. Второе важно, чтобы преждевременно не попасть в локальный оптимум, а получше «исследовать» всю часть пространства, отведённую для поиска.

Уже сейчас есть много механизмов для изменения баланса (они также описаны в 5.5.1).

Первый подход — «индивидуалистичный»: отдельное воздействие на те особи, которые «выделяются из серой массы», то есть всё ещё имеют относительно хорошие значения функции ошибки, но находятся в местах с низкой плотностью геномов. Под воздействием понимается увеличение `fitness`-функции или другие бонусы при размножении (например, «особая квота»).

Однако как уже существующие механизмы, так и предложение с отдельным воздействием действуют вслепую — по заранее заготовленному плану, не зависящему от текущей ситуации в популяции. Нужно перейти от задания действий перед запуском программы к заданию требуемых показателей — результатов действий, наладив работу подстраивающейся системы, регулирующей воздействие с помощью существующих механизмов.

А именно — нужно:

1. Задать метрику разнообразия, устойчивую к помехам и выдающую данные в человекочитаемом формате (чтобы задавать её зависимость).
2. На основе экспериментов определить, какой должна быть зависимость требуемого значения метрики от процента выполнения.
3. Научиться использовать регуляторные средства (5.5.1) для перевода популяции к заданному значению метрики.

8.5.1. Задание метрики

Важно, что она должна быть сравнима при подсчёте для разных функций и областей, желательно — находясь всё время в том же интервале, например, $\in [0; 1]$.

В простом случае можно считать метрику как отношение n -мерного «объёма» выпуклой оболочки к n -мерному «объёму» всего пространства поиска. (алгоритм нахождения МВО в n -мерном пространстве: https://neerc.ifmo.ru/wiki/index.php?title=%D0%92%D1%8B%D0%BF%D1%83%D0%BA%D0%BB%D0%B0%D1%8F_%D0%BE%D0%B1%D0%BE%D0%BB%D0%BE%D1%87%D0%BA%D0%B0_%D0%B2_n-%D0%BC%D0%B5%D1%80%D0%BD%D0%BE%D0%BC_%D0%BF%D1%80%D0%BE%D1%81%D1%82%D1%80%D0%BD%D1%81%D1%82%D0%B2%D0%B5)

Проблемой этой метрики может стать то, что в пространствах с высокой и низкой размерностью при таком же количестве точек могут быть несопоставимые результаты.

Однако понятно, что может быть несколько точек «по краям», а все остальные — сконцентрированы на одном пятаке. Метрика покажет, что точки достаточно диверсифицированы, что окажется неверным.

Варианты исправить этот недочёт такие: либо как-то отбросить некоторое количество процентов самых «далёких» точек ($\square \square$ «выбросы») или пытаться выделять «очаги» — скопления точек, либо производить те же операции учитывая все точки, но «нечётко»: например, вместо того, чтобы не учитывать точку, учитывать её с маленьким весом.

Проще всего — удалять каждый раз точку, максимально удалённую от «центра масс» (постоянно поддерживая актуальное его местоположение), убрав таким образом $\approx 10\%$ точек, а потом найти отношение объёмов.

Другой напрашивающийся подход — сначала посчитать в сетке точек (например, ≈ 10 по каждому измерению) значение плотности распределения геномов вокруг этой точки, потом для каждого генома определить, в зоне с какой примерной плотностью он находится и, например, просуммировать эти значения, потом сравнить результат со значением для идеально равномерного распределения и для всех геномов, находящихся в одной точке и откалибровать, переведя в шкалу от 0 до 1.

Плотность в точке оценивается по Гауссовым воротам, то есть:

$$density_{node} = \sum_{g \in genomes} \omega \left(\sqrt{\sum_{c \in coords} (g_c - node_c)^2} \right) \quad (14)$$

, где

$$\omega(d) = \frac{1}{\sigma \sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad (15)$$

Однако для пространств с высокой размерностью (то есть с большим количеством параметров) такое произвести не получится, так как количество узлов для

достаточной детализации становится слишком большим:

$$N_{nodes} = segments_per_axis^{dimensions} \quad (16)$$

То есть для 10-и делений по оси уже для 9-и параметров оказывается совершенно неприемлемое число операций.

Но в таком случае можно считать плотность не в узлах, которых становится больше, чем точек, находя каждый раз по требованию ближайший узел к точке, а в самих точках. Так мы перейдём от экспоненциальной сложности по отношению к количеству измерений к квадратичной по отношению к размеру популяции, что (учитывая её характерные значения), вообще не проблема. При магическом появлении большого количества вычислительных ресурсов и, соответственно, увеличения размера популяции можно для каждой точки учесть некоторую часть других, предположив, что распределение по расстоянию до неё примерно такое же, но вряд ли этот способ придётся применять.

Ещё вариант — считать разнообразие или плотность по каждой оси или комбинации осей отдельно, а потом — «просуммировать». Конечно, это не даёт исчерпывающей информации, но позволяет достаточно хорошо оценить разнообразие, а в вычислительном плане — несравненно быстрее.

Принципиально другой способ — использовать кластеризацию точек, и на основе её определять количество основных очагов точек.

8.5.2. Определение требуемой динамики разнообразия

Сначала динамика берётся «из головы», потом — подстраивается на основе экспериментов.

8.5.3. Регуляция разнообразности

Теперь у нас есть требуемая динамика некой величины и умение с разной «мощностью», как в одну сторону, так и в другую. Как нетрудно догадаться, в таких случаях имеет смысл использовать регуляторы, обеспечивающие отрицательную обратную связь, например, ПИД-регулятор:

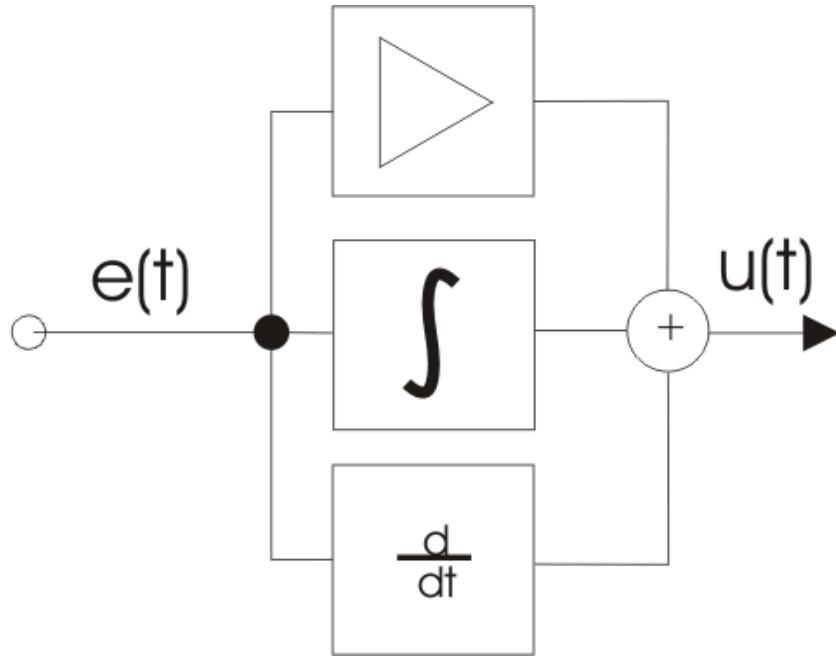


Рис. 20: Схема Пропорционально-Интегрально-Дифференциального регулятора

8.6. Улучшить алгоритм поиска цветов и разделения на зоны

Сейчас для разделения изображения на зоны используется Adobe Illustrator. По заданному количеству цветов (и, следовательно, уровню детализации) он разделяет изображение на зоны, присваивая каждой какой-то из цветов палитры так, чтобы он хорошо . Сама палитра тоже формируется в ходе работы алгоритма.

Скорее всего, для этого используется один из популярных алгоритмов, описанных здесь, или некая проприетарная их вариация. Зоны, на которые происходит деление, описываются частями плоскости, ограниченными кривыми безье — «path» формате *svg*.

Несмотря на то что формально алгоритм выполняет свою работу, большое количество зон имеет очень продолговатую форму, а также наблюдается неимоверный разброс в размерах между разными зонами (см. 7.1): всё это уменьшает эффективность процесса.

Я собираюсь написать свой алгоритм для этого, состоящий из примерно таких шагов:

- На основе распределения плотности точек по 3-мерному пространству цветов (возможно, стоит перевести в Lab color space) выделить заданное количество «островков» (представив точки изображения как график): «поднимать воду», пока количество не станет таким, но при реализации делать это, конечно, не линейным поиском.
- Когда островки определены — для каждого найти характеризующий его цвет.
- Начать относить пиксели к тому или иному цвету: начинать с тех, которые к какому-то из них очень близки, а затем выбирать цвет так, чтобы не создавать лишние зоны. Например, запустить из полеченных островков независимо обход в ширину.

- Возможно, ещё запустить склеивание соседних зон, учитывая их размер и разницу цветов
- Когда точно известно, какие точки получили цвет с каким id , имеет смысл обновить цвета для каждого id , чтобы они максимально соответствовали тем точкам, которые их «выбрали». Опять же — делать это можно через среднее арифметическое. Здесь важно сказать, что такой подход не дискредитирует идею большей детализации в районе часто встречающихся цветов и не превратит всё в кашу из близких цветов, так как, изначально эти близкие цвета были детализированы и, следовательно, выбирались соответствующие пиксели. Всё, что сделает последний шаг — немного подстроит результат, улучшив его.

8.7. Перенести графические вычисления на видеокарту

Также напрашивающееся улучшение. Это может существенно ускорить работу алгоритма, особенно — генетического (так как при нём можно распараллелить вычисление для целой популяции), причём только в случае, если не используется быстрый пересчёт функции ошибки или мутирует очень много мазков одновременно. Как бы то ни было, когда-нибудь стоит добавить эту возможность. Тестовый проект с использованием OpenCL я уже написал.

Изначальная картинка будет переноситься в видеопамять один раз — в начале работы программы. Выделить память для матрицы можно также один раз, а потом каждый раз её очищать (этую операцию можно производить параллельно).

Если использовать видеокарту для распараллеливания вычислений, встаёт вопрос, на каком именно уровне производить разделение. Варианты такие:

1. Каждое изображение — на своём потоке. Такой способ подходит только для ГА в случае огромного размера поколения (так как потоков у видеокарты порядка нескольких тысяч). Для отжига — никогда.
2. Каждый мазок — на своём потоке. Тут возникают проблемы с синхронизацией, так как порядок наложения важен: как минимум не должны появляться в хаотическом порядке пиксели из мазков разного цвета. Даже если происходит смешение цветов при наложении, синхронизация важна. Это можно сделать через дополнительную структуру данных в виде прямоугольной матрицы, в которой для каждого пикселя будет записываться список цветов с приоритетами (индексами мазков, а значит, и числами, определяющими порядок слоёв). Потом уже независимо для каждого пикселя будет происходить обработка смешения или наложения с замещением «попавших» на него цветов. Всё упирается в умение синхронизировать потоки. Тут нужно либо симуляцию мьютекса для каждого пикселя (то есть матрицу булевых значений «занят-не занят»), либо умение распределить мазки по потокам так, чтобы в каждый момент времени не было никаких двух с накладывающимися *bounding-box'ами*.

В первом случае либо нужно покупать видеокарту с поддержкой атомарных значений, либо придётся вставлять дополнительные проверки, чтобы два потока не прочитали почти одновременно состояние «не занят» и не зашли туда до того, как какой-либо из них успеет записать состояние «закрыто».

Во втором случае уменьшится количество потоков, одновременно работающих над одной «картиной».

3. Проводить распараллеливание на графическом примитиве низшего уровня, использующемся в данном алгоритме (см. 4.3). Например, полигоны или круги. Видеокарта умеет эффективно отрисовывать такие примитивы. Однако количество точек в примитивах обычно невелико: существенно меньше, чем ядер в видеокарте.

Причём всегда можно комбинировать разделения на разных уровнях.

Когда будет произведена растеризация, на той же видеокарте посчитается функция ошибки: в этом случае легко сделать это независимо для каждого пикселя. Единственное — нужно помнить, что для добавления наказания за наложения и пустоты в функцию ошибки надо составлять дополнительные матрицы, в которых это будет указано.

Адаптация алгоритмов под видеокарту описана здесь: 4.3.

8.8. Выделение границ мазками

Ни для кого не секрет, что существует много алгоритмов, позволяющих с неплохой точностью выделять «границы» у изображения — резкие цветовые (а иногда и содержательные, основанные на границах распознанных объектов или их паттернов!) переходы. Предположительно, их использование могло бы повысить качество работы алгоритма: дополнительная проработка границ увеличит чёткость и «читаемость» картины. Остается вопрос — как расположить точки, задающие мазки, чтобы робот обвёл границы объектов. Конечно же, будет применяться оптимизация, а «нарезать» линии границ можно, например, по их резким поворотам.