Получение предобученных данных для машинного обучения на базе процедур иерархической бинаризации и агрегации

Рихтер А.А.[а,б,в], Мурынин А.Б.[а,в,г], Филимонов А.А.[а], Харченко В.Д.[а]

[а] Научно-исследовательский институт аэрокосмического мониторинга "АЭРОКОСМОС", г. Москва

[б] Акционерное общество «Тазмар АйТи-солюшнз», г. Санкт-Петербург

[в] Государственный университет по землеустройству, Москва, РФ

[г] ФИЦ ИУ РАН, Москва, Россия

urfin17@yandex.ru

Постановка задачи

- В работе представлен способ подготовки данных для машинного обучения для семантической сегментации информативных классов на изображениях, основанный на построении моделей кластеризации.
- Предлагается метод ускорения подготовки обучающей выборки на основе кластеризации областей на изображениях и векторизации границ.
- В работе предлагается способ ускорения подготовки данных для семантической сегментации информативных классов на изображениях. Данный способ основан на получении модели кластеризации, построенной на базе применения операций иерархической бинаризации и агрегации.

Методы кластеризации

Одним из наиболее популярных методов кластеризации является k-means.

- Преимущества метода: 1) прост в реализации и понимании; 2) высокая скорость работы и точность на данных сферической формы; 3) 2) наличие большого числа модификаций.
- Недостатки метода: 1) задание числа кластеров / их начальных точек кластеров; 2) сходится к локальным максимумам, что даёт несколько разные результаты кластеризации каждый раз при постоянном числе кластеров; 3) кластеризуются сфероподобные области (т.е. эллипсо- или линейно подобные области кластеризуются значительно хуже); 4) в целом не учитывает плотность данных и неоднородность кластера.

Разные методы кластеризации могут иметь иерархические расширения. Основными параметрами иерархических методов являются метод вычисления связи и метрика связи.

- Преимущества методов: 1) способность обнаружения кластеров произвольной формы; 2) работоспособность с различными паттернами данных; 3) возможность формирования информативной иерархии кластеров для лучшего понимания структуры данных; 4) возможность получения оптимальной кластеризации; 5) производит значительно ниже шума.
- Недостатки методов: 1) использование большого количества вычислительных ресурсов и памяти из-за работы со всей матрицей расстояний между объектами; 2) чувствительность к выбору критерия объединения кластеров и неустойчивость к шуму и выбросам, что может сильно искажать иерархию кластеров.

Общая схема кластеризации изображений

Общая схема кластеризации изображений

- Кластеризацию можно представить отображением $f(p, I) \rightarrow K$, где I – изображение для кластеризации, p – параметры кластеризации, f – метод и алгоритм кластеризации, K оценённая маска кластеров (бинарная, пообъектная или маска полей энергии), размер которой равен размеру I.
- Итерируемые параметры $p' \subset p$ – параметры кластеризации, значения которых меняется для обеспечения оптимизации некоторой целевой функции.
- Основными параметрами кластеризации, которые также относятся к итерируемым параметрам, являются: количество m кластеров $\{k_i\}$, метрика схожести / разделимости кластеров μ. В качестве μ могут быть метрики расстояний, угловые метрики, метрики на основе корреляций и др. [10-11]
- Для оптимизации может быть подана эталонная маска кластеров $K_Э$ (эталоны объектов) изображения I. Оптимизация производится на базе целевой функции $F(p', \mu, K, K_Э)$ или её частного случая $F(p', \mu, K)$ при отсутствии эталонной кластеризации. Значение p', при котором F достигает экстремума (минимального или максимального), считается оптимальным значением, при всех неизменных других p\p' параметрах кластеризации.

Тайлирование

- Тайлирование (тайлинг) – разделение данных на части (тайлы), каждая из которых обрабатывается некоторым алгоритмом в отдельности. Как правило, необходимость в тайлировании обусловлена наличием больших размеров данных. При этом обработка алгоритмом данных целиком выражается через обработку алгоритмом тайлов данных в отдельности. Для обеспечения эквивалентности такой обработки необходимо провести обратную процедуру, то есть слияние обработанных тайлов.
- Формально тайлирование можно записать в виде:
- $f(I) = \left\{ \mathrm{U}_{j=1}^{l} f(I_j), \bar{f}\left(\{I_j\}\right) \right\}, I = \mathrm{U}_{j=1}^{l} I_j.$
- где f – алгоритм обработки; I – данные целиком; I_j, j=1…l – тайлы разбиения данных; l – количество тайлов разбиения; $\bar{f}$ – алгоритм слияния тайлов, порождаемый алгоритмом f. Над разбиением $\{I_j\}$ данных производится $\bar{f}$ такой, чтобы совместно с процедурами $f(X_1)$, …, $f(I_j)$ результат обработки был равен или приближённо равен результату обработки процедуры f(I).

Примеры схем тайлирования

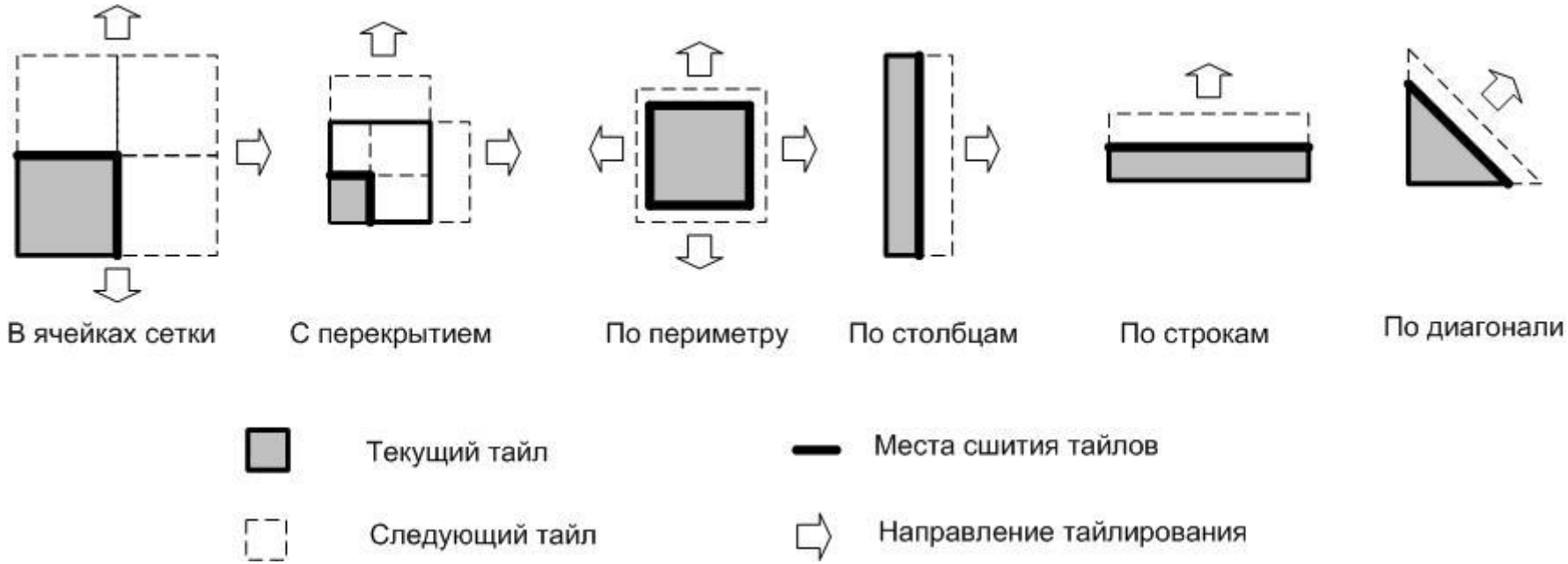
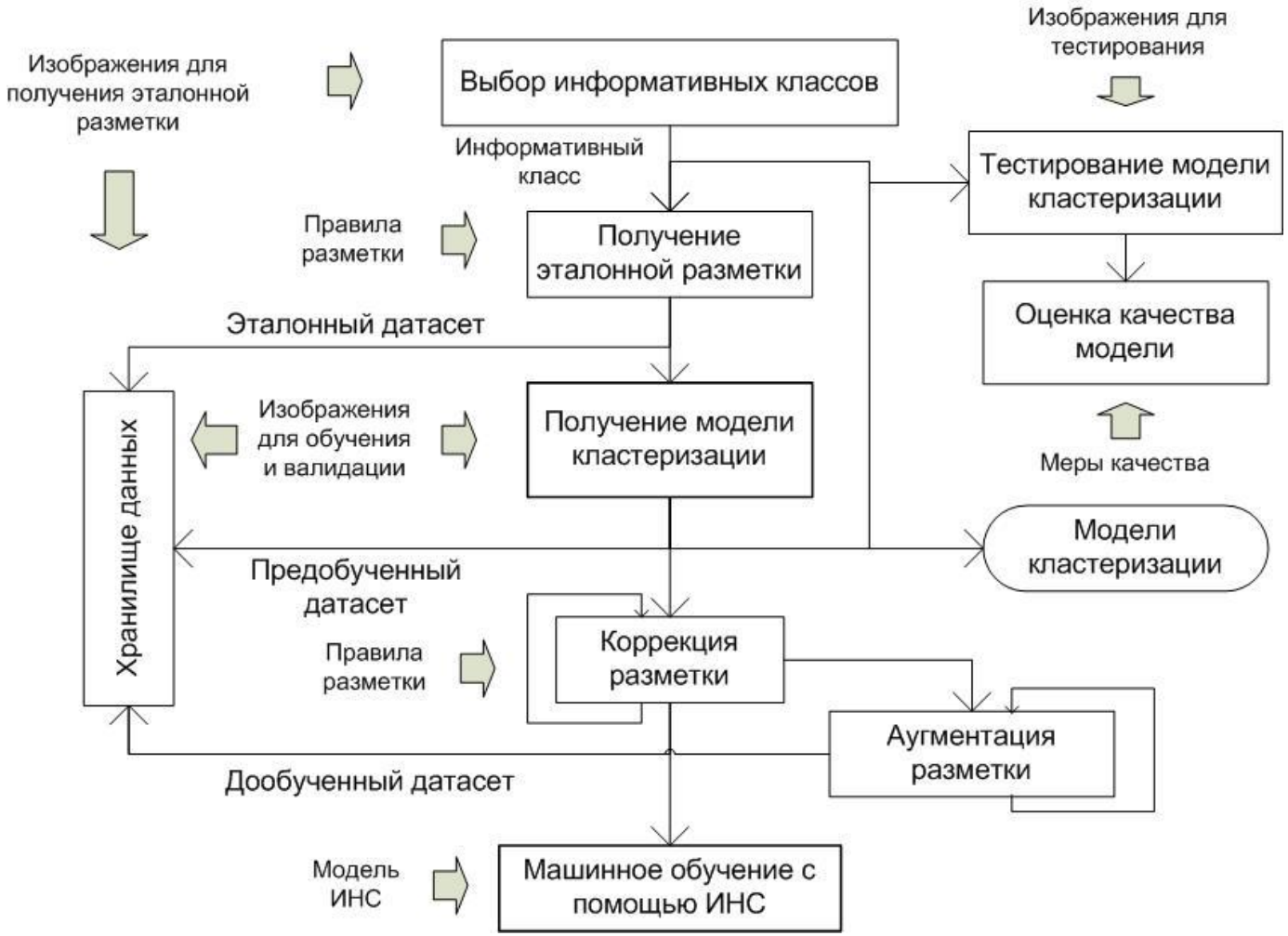


Схема подготовки данных для машинного обучения



Эталонный датасет – формируется для каждого информативного класса «ручным» и интерактивным способом

Предобученный датасет – формируется в результате расчёта модели кластеризации

Дообученный датасет – формируется по результатам коррекции и аугментации предобученной, выполняемые на кластеризованных областях.

Схема подготовки данных для машинного обучения

Аугментация обучаемых данных позволяет дополнительно «умножить» разметку с применением различных видов аугментации для размеченных областей. В частности: 1) изменение оттенка, насыщенности или яркости; 2) повороты, масштабирования или сдвиги; 3) преобразования перспективы; 4) копирование областей для стационарных объектов (имеющих постоянную локацию на земной поверхности) на изображения этих объектов в другие моменты времени с тем же ракурсом съёмки.

В хранилище данных содержится: 1) изображения и их метаданные (для эталонной разметки, обучаемые, валидируемые и тестируемые при кластеризации); 2) модели кластеризации каждого информативного класса; 3) маски эталонной разметки и обучающей выборки информативных классов.

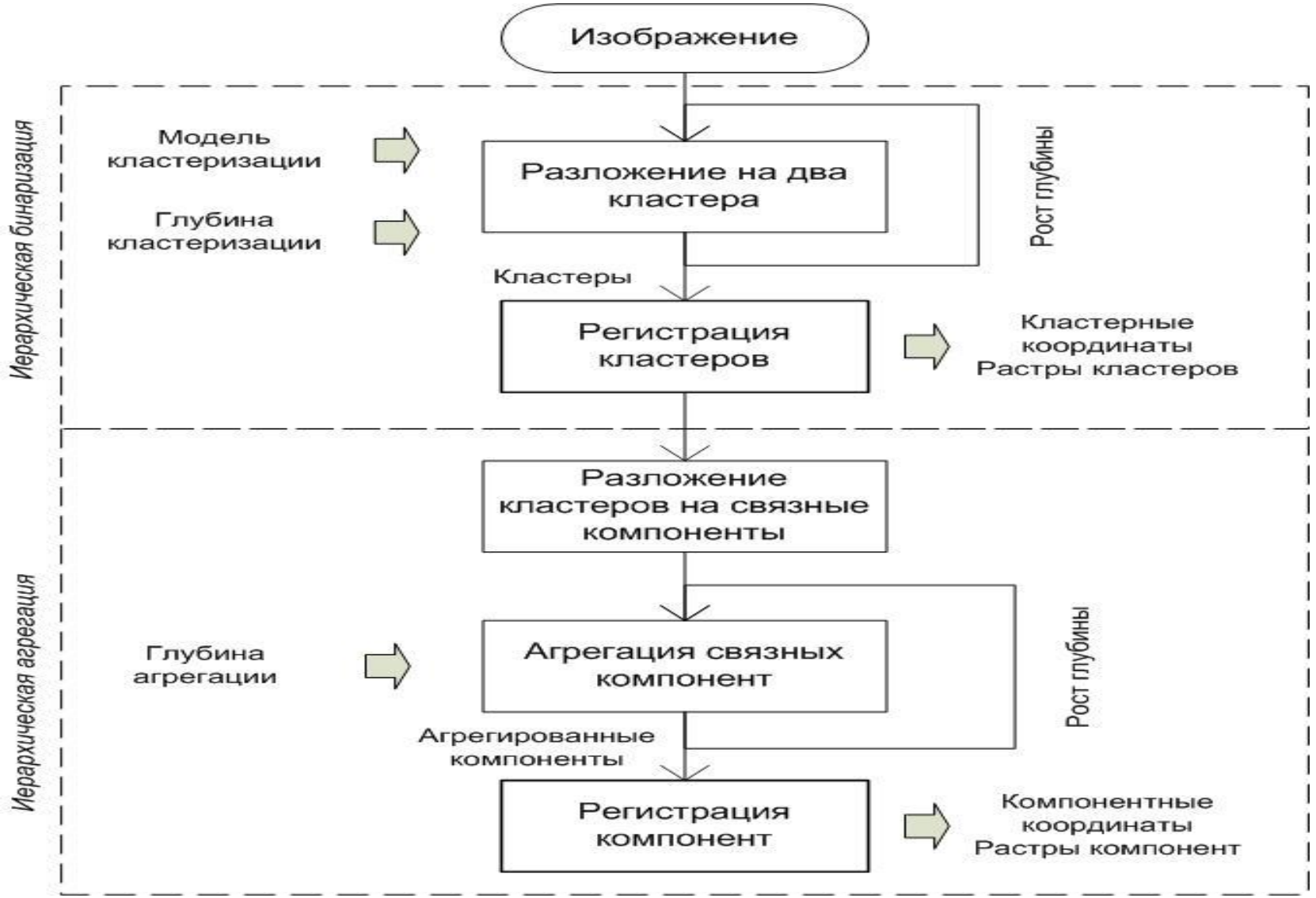
На вход модели нейронной сети подаётся обучающая выборка для эталонной разметки, а также разметки, рассчитанной посредством кластеризации и прошедшей процедуры коррекции и аугментации.

В качестве модели сети берётся одна из ранее разработанных моделей свёрточных сетей для сегментации экологических объектов [2]. Это многоклассовые и бинарные модели для сегментации различных экологических классов – мусорных свалок, разных типов жидкостей, объектов жидкостной и трубопроводной инфраструктуры, объектов автодорожной и железнодорожной инфраструктуры, зданий (в том числе производственных). Модели обучены на полутоновых и мультиспектральных изображениях импактных районов Арктики и Московского региона. Данные аугментированы с применением процедуры ROI Cover. В качестве функции потерь использовалась функция потерь Жаккарда (Joccard loss), а также данная функция в комбинации с бинарной перекрестной энтропией.

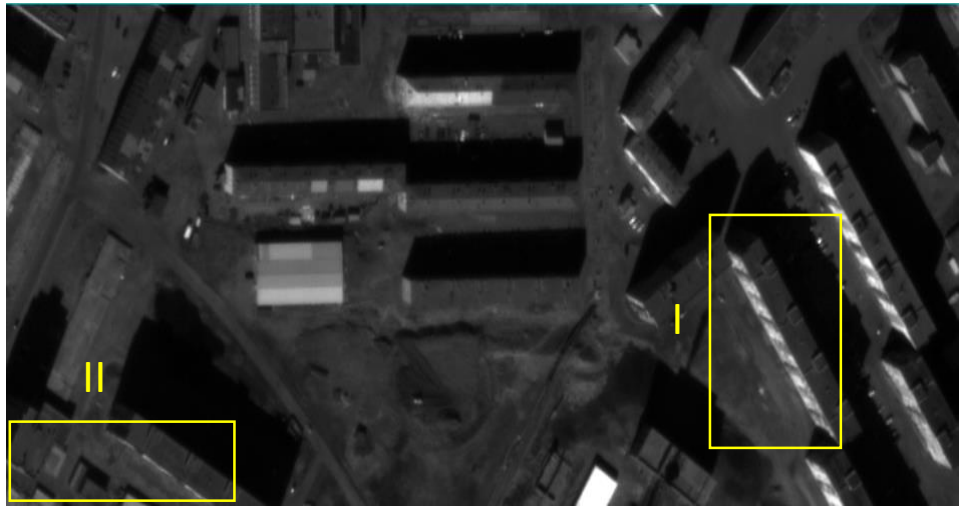
Типы правил построения обучающей выборки

Тип правила	Описание	Пример правила для информативного класса
Правила локализации	Обнаружение рабочей области поиска объектов	Свалки сосредотачиваются с большей вероятностью на периферии города, чем в его центре
Правила обнаружения	Обнаружение объектов в рабочей области на базе геометрических, текстурных, яркостных и др. признаков	Одним из геометрических признаков окон на стене здания является регулярность их распределения
Правила выделения	Разграничение объекта от фона	Граница между мусорным захламлением и фоном лежит в местах максимальных перепадов зернистости текстуры захламления
Правила векторных типов	Особенности использования векторов при разметке	Ребро здания выделяется линейным вектором некоторой «толщины» (линейный вектор автоматически модифицируется в площадной с данной толщиной)
Правила предположения	Как размечать в условиях невидимости части объекта	В местах небольшой протяжённости заслонения дороги деревьями или тенями предполагается такая же дорога
Правила разделимости	Как размечать в условиях наложения одних объектов на другие	Вагоны подвижного состава размечаются непересекающимися областями
Правила артефактов	Как размечать в условиях артефактов съёмки	При наличии геометрических искажений крыши здания данная область исключается из обучающей выборки
Правила интерпретации	Разметка в условиях неоднозначности или сложности в интерпретации	При схожести крыши здания с объектом из другого класса данная область исключается из обучающей выборки
Правила унификации объекта	Разметка в условиях типизации объектов	Здание и его крыши имеет уникальную или сложную форму
Правила исключения разметки	Как исключается область из обучения	Возможность построения исключаемых областей полигонами
Правила коррекции разметки	Как достраивать, удалять, перемещать разметку / часть разметки объекта	В местах наличия ложной разметки на части объекта (ошибки первого рода) построение стираемых областей полигонами

Процедуры иерархической бинаризации и агрегации

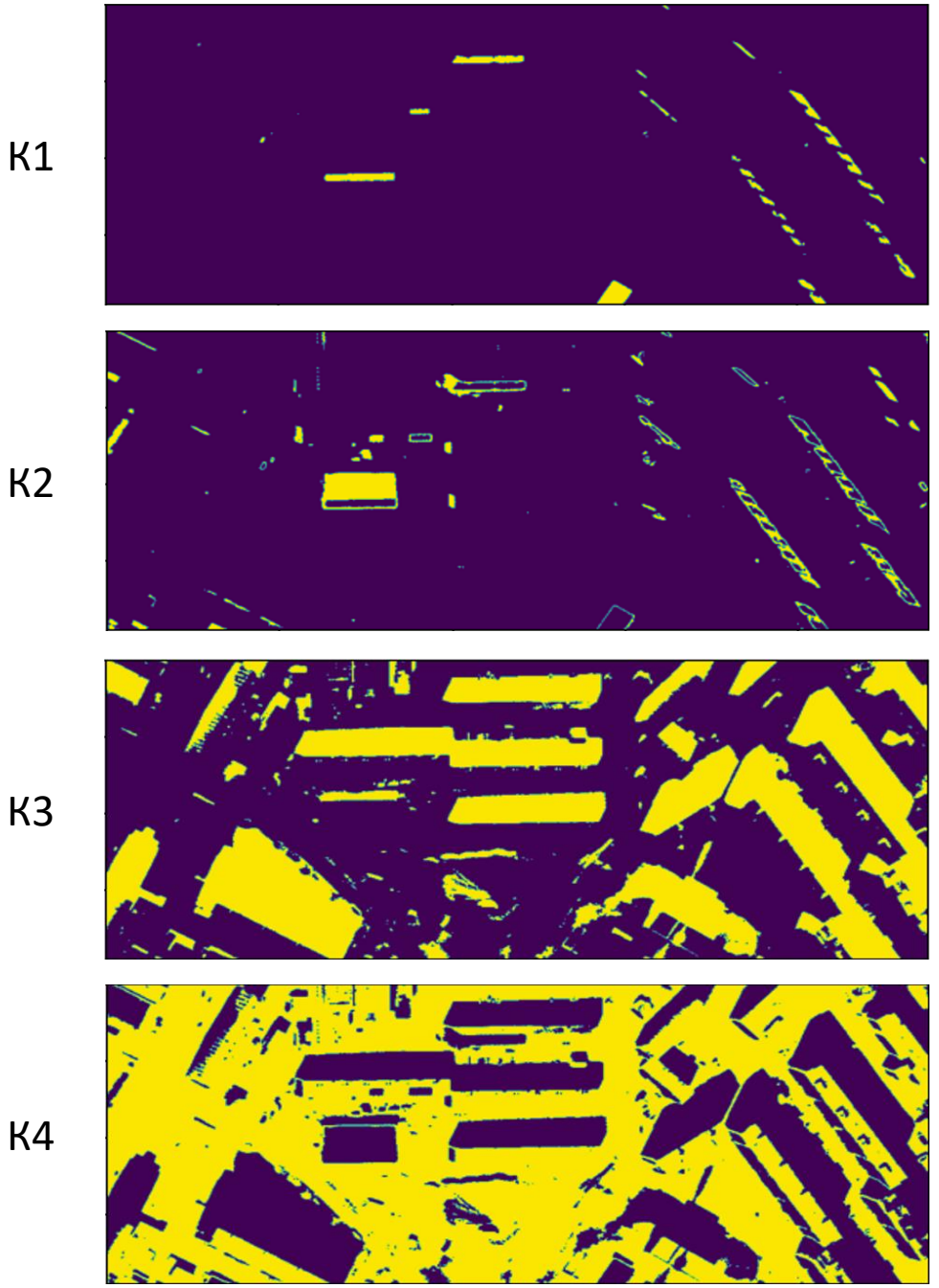


Пример иерархической бинаризации

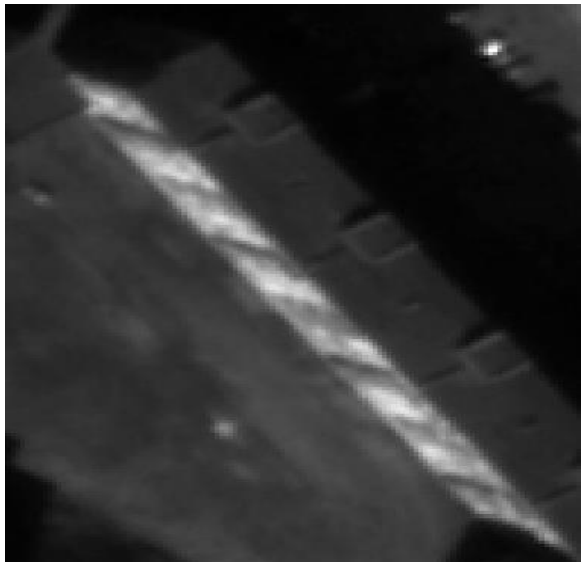


Входное изображение

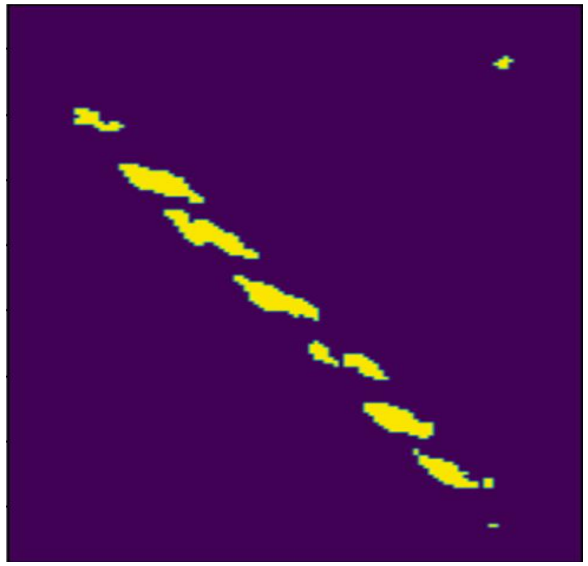
Кластер	Координаты кластера	Физический смысл
К1	00	Освещённые части здания (крыши и стены)
К2	01	Полуосвещённые части здания (крыши и стены)
К3	10	Тени от зданий
К4	11	Прочее

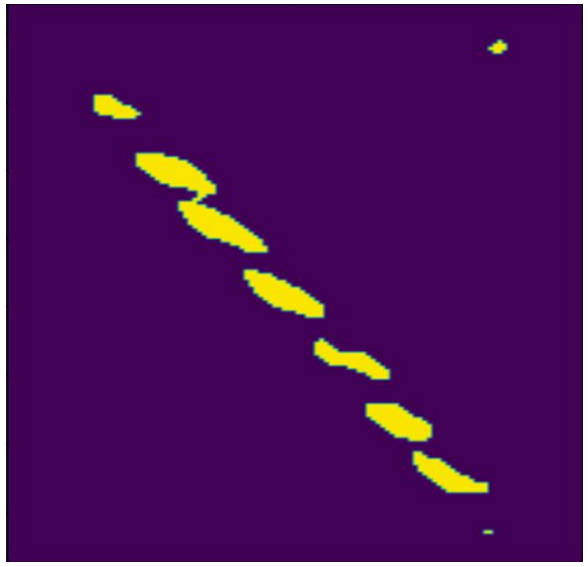


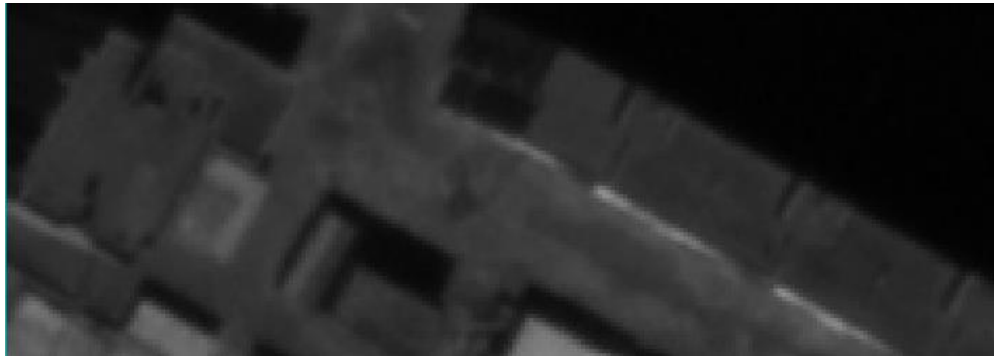
Пример иерархической агрегации



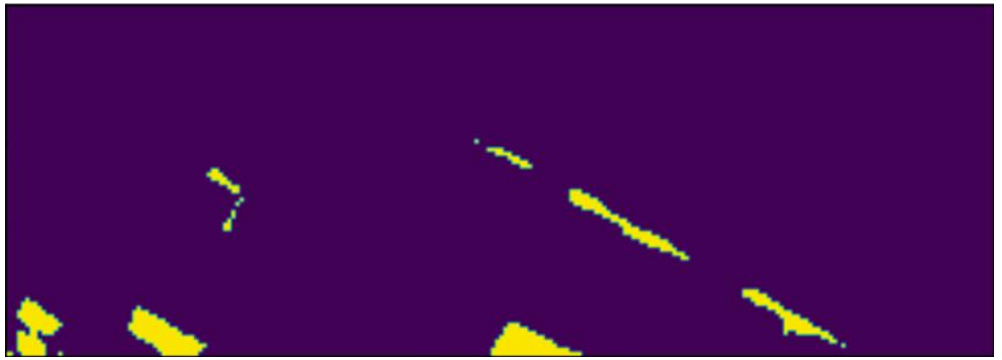
Фрагмент I

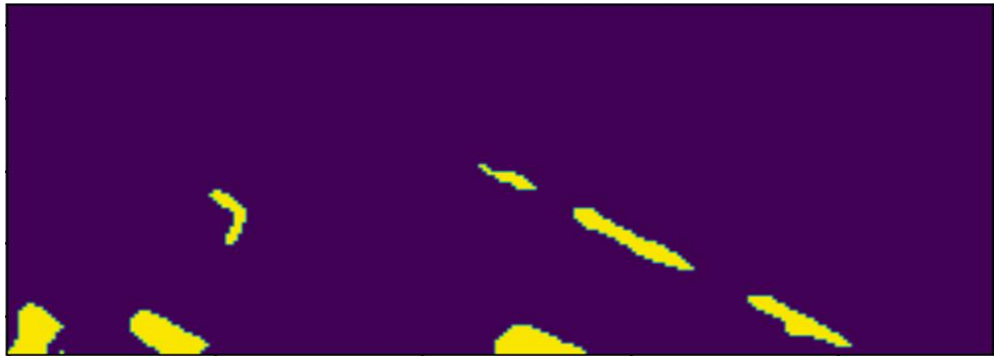






Фрагмент II





База данных спутниковых систем ДЗЗ

Разработана база данных спутниковых систем дистанционного зондирования Земли, применяемая для выбора категории спутниковых изображений, релевантных требованиям (пространственного и спектрального разрешения, временных и геометрических характеристик и др.) для получения приемлемого качества модели кластеризации.
База данных спутников дистанционного зондирования Земли содержит только активные спутники или спутники, планируемые к запуску. Полностью актуализируется раз в 3 месяца.

Разделена на 4 блока:

1. «Система» - короткая информация о спутнике (название, статус, дата запуска, стартовая площадка и т.д., а также полетные характеристики (апогей, перигей, наклонение и т.д.);
2. «Аппаратура» - техническая информация о приборе, стоящем на спутнике (пространственное разрешение, полоса обзора, спектральный диапазон и др.);
3. «Принадлежность» - какая страна, кто оператор/конструктор и какого типа спутник;
4. «Запуск» - информация о месте запуска.

Выводы

- В работе предлагается способ ускорения подготовки данных для семантической сегментации информативных классов на изображениях. Данный способ основан на получении модели кластеризации, построенной на базе применения операций иерархической бинаризации и агрегации.
- При иерархической бинаризации каждый кластер рекурсивно раскладывается на два кластера с применением неиерархического метода кластеризации. При иерархической агрегации происходит разложение каждого кластера на связные компоненты и рекурсивно объединение компонент, расстояние между которыми не более растущего порогового значения. Каждый объект модели координируется, в зависимости от числа рекурсивных обращений при бинаризации и агрегации.
- Данный способ автоматизации получения первичных информативных признаков является составной частью процедуры подготовки данных для машинного обучения.
- Данный способ автоматизации получения первичных информативных признаков является составной частью процедуры подготовки данных для машинного обучения.
- В отличие от большинства существующих иерархических методов данный упрощённый подход не требует больших затрат потребляемой памяти и позволяет его применять при тайловой обработке, при обработке больших изображений, мультиспектральных и гиперспектральных изображений.
- Разработанная база данных спутниковых систем дистанционного зондирования, применяемая для получения приемлемого качества модели кластеризации, может стать ценным инструментом для быстрого поиска информации о конкретном космическом аппарате с возможностью создания приложений и программного обеспечения для их отслеживания.