

© А. С. Фатеев, Д. П. Кричевец, Е. Л. Киселев,
Д. А. Гаврилов, А. А. Фортунатов, Е. А. Татарина, 2026

СТРУКТУРНО-ОРИЕНТИРОВАННЫЙ ПРЕДВАРИТЕЛЬНЫЙ ОТБОР ВИДЕОКАДРОВ БЕСПИЛОТНОГО ЛЕТАТЕЛЬНОГО АППАРАТА ДЛЯ ЗАДАЧ КРОСС-МОДАЛЬНОГО СОПОСТАВЛЕНИЯ СО СПУТНИКОВЫМИ ИЗОБРАЖЕНИЯМИ

Рассматривается задача предварительного отбора наиболее информативных видеок кадров для системы визуальной навигации беспилотных летательных аппаратов, основанной на кросс-модальном сопоставлении со спутниковыми изображениями в условиях ограниченных вычислительных ресурсов. Высокая частота формирования видеопотока приводит к существенному объему входных данных, тогда как модели сопоставления обладают ограниченной пропускной способностью. В результате в режиме реального времени обрабатывается лишь часть кадров, что повышает риск пропуска наиболее информативных изображений. Показано, что кадры, содержащие геометрически устойчивые структуры сцены, наблюдаемые на разных масштабах изображения и имеющие согласованную ориентацию границ, обладают более высокой вероятностью успешного сопоставления со спутниковыми изображениями. Обоснована необходимость использования структурных критериев, ориентированных на задачи сопоставления.

Ключевые слова: беспилотные летательные аппараты, кросс-модальное сопоставление изображений, предварительный отбор кадров, структурная информативность, вейвлет-преобразование, мультимасштабный анализ, спутниковые изображения, визуальная навигация

ВВЕДЕНИЕ

Кросс-модальное сопоставление изображений, направленное на установление пространственных соответствий между изображениями, полученными различными сенсорными системами и при существенно отличающихся условиях съемки, является одной из ключевых задач компьютерного зрения. В контексте дистанционного зондирования и автономных систем данная задача приобретает особую значимость, поскольку точное сопоставление изображений служит основой для географической привязки, навигации, реконструкции сцены и принятия решений в реальном времени.

В частности, сопоставление изображений, полученных с беспилотных летательных аппаратов (БПЛА) и спутников, представляет собой крайне сложную кросс-модальную задачу. Изображения с камеры БПЛА и спутников принципиально различаются по масштабу, ракурсу наблюдения, пространственному разрешению и геометрической структуре сцены. Спутниковые изображения, как правило, формируются в ортогональной проекции и охватывают обширные территории, тогда как изображения БПЛА характеризуются ограничен-

ным полем зрения, выраженными перспективными искажениями и высокой вариативностью ракурсов. Эти различия существенно усложняют установление устойчивых соответствий между изображениями.

Современные системы БПЛА [1] оснащаются камерами с высокой частотой кадров, что приводит к формированию плотных видеопоследовательностей и значительных потоков визуальных данных. В то же время методы кросс-модального сопоставления изображений с камеры и спутника, как правило, основаны на вычислительно сложных моделях, включая глубокие нейронные сети, плотные [2] или полуплотные [3] алгоритмы сопоставления и процедуры геометрической нормализации. В результате пропускная способность таких моделей оказывается существенно ниже скорости поступления данных от сенсоров, что делает обработку всех кадров в режиме реального времени практически невозможной. В условиях ограниченных вычислительных ресурсов система вынуждена выполнять частичную обработку входного потока, что повышает вероятность пропуска кадров, обладающих наибольшей информативностью и обеспечивающих оптимальные условия для кросс-модального сопоставления.

В таких условиях особую важность приобретает задача предварительного отбора кадров, целью которой является сокращение входного потока данных за счет исключения кадров, малопригодных для последующего кросс-модального сопоставления. В отличие от классических задач видеоселекции, здесь требуется отбор не просто визуально качественных изображений, а кадров, обладающих высокой информативностью именно с точки зрения успешного сопоставления со спутниковыми изображениями.

Наиболее распространенные подходы к отбору кадров основываются на использовании общих метрик качества изображения, таких как резкость, контрастность или энергия в частотной области [4–6]. Подобные критерии эффективно выявляют технически дефектные кадры, например, размытые или зашумленные изображения. Однако в задачах кросс-модального сопоставления такие метрики оказываются недостаточными. Визуально четкие и контрастные кадры БПЛА не гарантируют наличия устойчивых структурных признаков, согласованных с геометрией спутниковых изображений, и, следовательно, не обязательно приводят к успешному сопоставлению [6, 7].

Основной причиной этого является то, что классические метрики качества оценивают преимущественно низкоуровневые визуальные характеристики, не учитывая структурную и геометрическую совместимость изображений различных модальностей. В частности, мелкие текстурные детали, тени, подвижные объекты и локальные изменения освещения могут существенно повышать визуальное "качество" кадра, но при этом не вносить полезной информации для сопоставления со спутниковыми изображениями, в которых доминируют крупномасштабные и стабильные геометрические структуры, такие как дорожная сеть, застройка или границы ландшафта.

Таким образом, возникает необходимость в разработке методов предварительного отбора кадров, ориентированных не на субъективное или визуальное качество изображения, а на оценку его потенциальной пригодности для кросс-модального сопоставления. Такие методы должны учитывать структурную информативность кадров, их устойчивость к изменениям масштаба и ракурса, а также способность сохранять согласованную пространственную организацию сцены между изображениями БПЛА и спутника.

В данной работе предлагается подход к предварительному отбору кадров с камеры БПЛА, направленный на повышение эффективности кросс-модального сопоставления со спутниковыми изображениями в условиях ограниченных вычислительных ресурсов [8, 9]. В отличие от существующих методов, ориентированных на общие метрики

качества изображения, предлагаемый метод фокусируется на анализе структурных характеристик изображения, релевантных задаче сопоставления. Это позволяет существенно сократить количество обрабатываемых кадров без потери точности географической привязки и повысить практическую применимость систем кросс-модального сопоставления в реальных сценариях.

ПОСТАНОВКА ЗАДАЧИ

Пусть $F = I_{t(t=1)}^T$ — видеопоследовательность, полученная с бортовой камеры беспилотного летательного аппарата с высокой частотой кадров, где T обозначает общее число кадров в рассматриваемом временном интервале, а I_s — спутниковое изображение соответствующей территории.

Ввиду ограниченных вычислительных ресурсов обработки всех T кадров видеопоследовательности для кросс-модального сопоставления является практически невозможной. В связи с этим требуется выбрать подмножество кадров

$$F' \subset F, |F'| \ll T,$$

которые обладают наибольшей пригодностью для последующего сопоставления изображений БПЛА со спутниковыми данными.

В отличие от традиционных методов отбора кадров, основанных на визуальных метриках качества изображения, в данной работе ставится задача оценки структурной информативности кадра, отражающей его способность обеспечивать устойчивые геометрические соответствия в задаче кросс-модального сопоставления.

МУЛЬТИМАСШТАБНЫЙ СТРУКТУРНЫЙ АНАЛИЗ С ИСПОЛЬЗОВАНИЕМ 2D-ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ

Описание преобразования

Для анализа структурных характеристик изображения на различных пространственных масштабах используется двумерное дискретное вейвлет-преобразование [10, 11], обеспечивающее совместную пространственно-частотную локализацию признаков.

Пусть $I \in R^{H \times W}$ — входное изображение. Двумерное дискретное вейвлет-преобразование строится на основе последовательного применения одномерных фильтров по строкам и столбцам изображения. Пусть g и h обозначают соответственно высокочастотный и низкочастотный фильтры анализа. Тогда операции вейвлет-разложения могут быть записаны в виде

$$L = (I * h) \downarrow 2, \quad G = (I * g) \downarrow 2,$$

где $(*)$ обозначает свертку, а $\downarrow 2$ — операцию двукратной дискретизации (downsampling).

После последовательной фильтрации по строкам и столбцам формируются четыре поддиапазона:

$$(A_l, H_l, V_l, D_l),$$

где A_l — аппроксимирующий поддиапазон (низкие частоты по обоим направлениям), а H_l , V_l и D_l представляют соответственно горизонтальные, вертикальные и диагональные детали изображения.

Аппроксимирующий поддиапазон A_l рекурсивно используется для построения следующего уровня разложения, что приводит к иерархическому мультимасштабному представлению изображения:

$$I \rightarrow \{A_L, H_L, V_L, D_L, \dots, H_1, V_1, D_1\},$$

где L обозначает число уровней разложения.

Такое представление позволяет выделять структурные компоненты изображения на различных пространственных масштабах, что делает вейвлет-разложение удобным инструментом для анализа геометрически устойчивых элементов сцены, таких как контуры, текстурные паттерны и направленные структуры. Пример вейвлет-разложения синтетического изображения представлен на рис. 1. В качестве входных данных используется искусственно сформированная сцена, содержащая горизонтальные, вертикальные и диагональные линейные структуры. Применение синтетического изображения позволяет наглядно продемонстрировать направленную селективность 2D-вейвлет-преобразования в условиях контролируемой геометрии сцены, исключая влияние текстурных и фотометрических факторов, характерных для реальных данных.

Как видно из рис. 1, коэффициенты детализации первого уровня разложения выделяют структуры в соответствии с их ориентацией: горизонтальные компоненты концентрируются в поддиапазоне H_L , вертикальные — в V_L , а диагональные — в D_L . При этом аппроксимирующий компонент сохраняет общую геометрию сцены. Данный пример иллюстрирует способность вейвлет-преобразования выполнять пространственно локализованный и направленно-чувствительный анализ изображения, что является ключевым свойством при оценке структурной информативности кадров для сопоставления.

В отличие от преобразования Фурье [12], обладающего глобальной частотной локализацией, вейвлет-преобразование сохраняет пространственную локализацию структурных элементов. Это особенно важно для изображений, полученных с камеры БПЛА, которые часто содержат выраженные границы объектов, линейные структуры (дороги, здания, границы полей) и текстурные элементы различного пространственного масштаба. В ряде работ для оценки информативности изображений используются более простые локальные показатели, такие как энтропия изображения [13], градиентные метрики [14], мера резкости на основе Лапласиана [15], а также характеристики, основанные на тензоре структуры [16]. Данные методы отражают отдельные аспекты локальной контрастности или статистического разнообразия изображения, однако не учитывают взаимосвязь структурных элементов на различных пространственных масштабах. Использование вейвлет-разложения позволяет анализировать распределение энергии структурных компонент одновременно в пространственной и частотной областях, что делает данный подход более подходящим для выявления геометрически устойчивых структур сцены, значимых для задач установления межизображенческих соответствий.

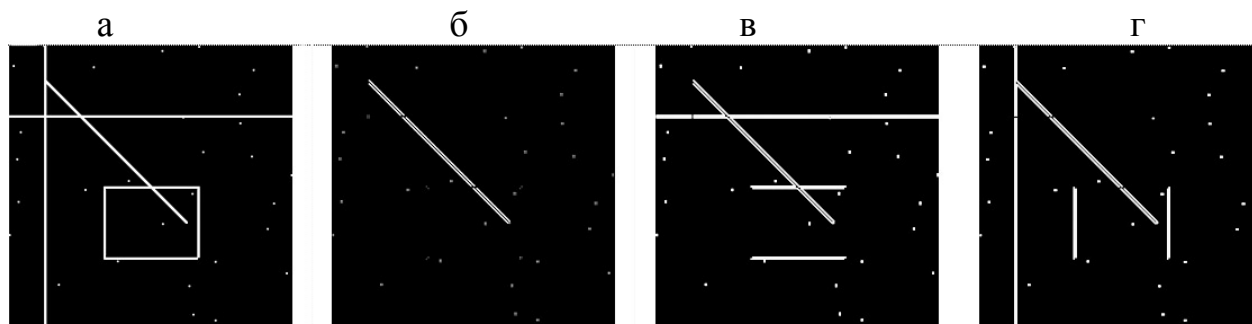


Рис. 1. Пример 2D-вейвлет-разложения синтетического изображения.

а — исходное изображение; б — диагональные коэффициенты детализации D_1 первого уровня; в — горизонтальные коэффициенты H_1 первого уровня; г — вертикальные коэффициенты V_1 первого уровня

Оценка структурной информативности в вейвлет-домене

Пусть для изображения I выполнено L -уровневое 2D-вейвлет-разложение, в результате которого получены аппроксимирующие и детализирующие поддиапазоны

$$(A_l, H_l, V_l, D_l), \quad l=1, \dots, L.$$

Поддиапазоны H_l , V_l и D_l представляют собой матрицы коэффициентов детализации на уровне l , соответствующие горизонтальным, вертикальным и диагональным структурам изображения.

Предполагается, что геометрически устойчивые структуры сцены проявляются в виде локальных максимумов энергии в детализирующих поддиапазонах на нескольких масштабах. Для количественной оценки вводится нормированная энергетическая характеристика уровня l :

$$E_l = \frac{1}{|H_l|} \sum_{i \in H_l} H_l(i)^2 + \frac{1}{|V_l|} \sum_{i \in V_l} V_l(i)^2 + \frac{1}{|D_l|} \sum_{i \in D_l} D_l(i)^2,$$

где $H_l(i)$ обозначает i -й коэффициент соответствующего поддиапазона, а модульные скобки $|\dots|$ — число коэффициентов в поддиапазоне.

Использование квадрата коэффициентов соответствует оценке энергии сигнала нормы L_2 :

$$\|H_l\|_2^2 = \sum_i H_l(i)^2,$$

что согласуется с энергетической интерпретацией вейвлет-преобразования как ортонормированного разложения. При использовании ортонормальных вейвлетов выполняется сохранение энергии, что позволяет интерпретировать E_l как долю структурной энергии изображения, локализованной на масштабе l .

Нормировка на число коэффициентов $|H_l|$, $|V_l|$, $|D_l|$ обеспечивает инвариантность оценки к размеру поддиапазона и позволяет корректно сравнивать уровни разложения различной размерности.

Таким образом, величина E_l отражает суммарную интенсивность направленных границ и протяженных структур на масштабе l , подавляя вклад случайного шума и мелкомасштабных текстур за счет квадратичного усреднения.

Межмасштабная согласованность структур

Для повышения устойчивости оценки структурной информативности вводится мера

межмасштабной согласованности, отражающая степень сохранения энергетических характеристик изображения при переходе между соседними масштабами разложения.

Для уровней l и $l+1$ межмасштабная согласованность определяется как

$$C_l = 1 - \frac{|E_l - E_{l+1}|}{E_l + E_{l+1} + \varepsilon},$$

где $\varepsilon > 0$ — малый регуляризирующий параметр, обеспечивающий численную устойчивость при $E_l = E_{l+1} = 0$.

Величина C_l принимает значения в диапазоне $0 \leq C_l \leq 1$ и характеризует степень сходства распределения энергии на соседних масштабах. Значения C_l , близкие к единице, соответствуют случаям, когда энергетические характеристики практически не изменяются при переходе между масштабами, что свидетельствует о наличии геометрически устойчивых и протяженных структур сцены.

Напротив, малые значения C_l указывают на существенные различия между E_l и E_{l+1} , что характерно для мелкомасштабных текстур, шума или локальных артефактов, энергия которых быстро затухает при укрупнении масштаба.

Функция оценки и стратегия отбора кадров

Для обеспечения сопоставимости вкладов различных масштабов энергетические характеристики нормируются по полной энергии детализирующих поддиапазонов:

$$\tilde{E}_l = \frac{E_l}{\sum_{k=1}^L E_k + \varepsilon},$$

где $\varepsilon > 0$ — малый регуляризирующий параметр, обеспечивающий численную устойчивость при малых значениях энергии.

Такая нормировка устраняет зависимость оценки от общей яркости изображения и обеспечивает выполнение условия

$$\sum_{l=1}^L \tilde{E}_l \approx 1.$$

Итоговая оценка структурной информативности кадра I определяется как взвешенная комбинация нормированных энергетических и межмасштабных характеристик:

$$S(I) = \sum_{l=1}^L \alpha_l \tilde{E}_l + \sum_{l=1}^{L-1} \beta_l C_l,$$

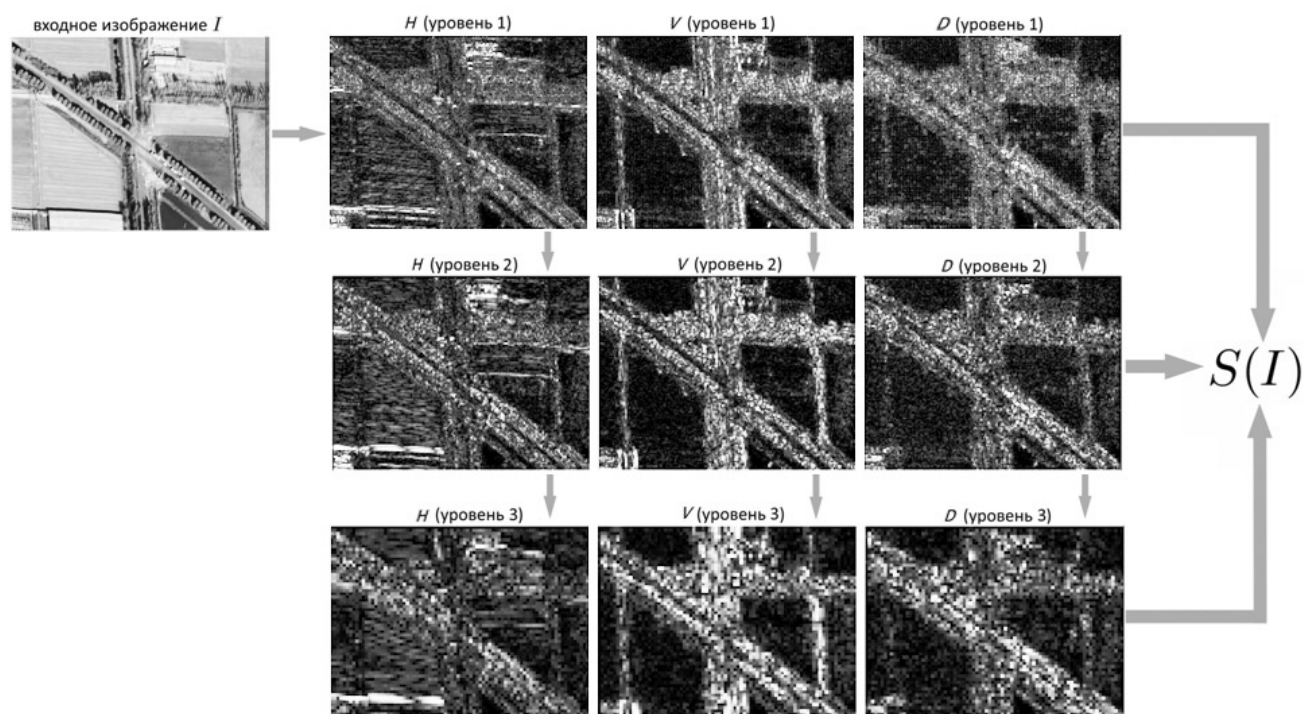


Рис. 2. Схема вычисления оценки структурной информативности кадра $S(I)$ на основе трехуровневого вейвлет-разложения изображения

где коэффициенты α_l и β_l задают вклад соответствующих масштабов и выбираются эмпирически. Первая сумма отражает распределение структурной энергии по масштабам разложения, а вторая — степень межмасштабной устойчивости выявленных структур.

Кадры видеопоследовательности ранжируются по значению $S(I_i)$, после чего формируется подмножество

$$F = \{I_i : S(I_i) \text{ входит в топ } K\},$$

где K — заданное число отбираемых кадров $\ll T$.

Упрощенная схема вычисления предложенной оценки представлена на рис. 2. На вход метода подается изображение перекрестка, снятого с камеры БПЛА, которое подвергается многоуровневому вейвлет-разложению. Для каждого уровня вычисляются энергетические характеристики детализирующих коэффициентов E_l и показатели межмасштабной согласованности C_l . Полученные величины агрегируются в итоговую оценку структурной информативности $S(I)$, которая используется для ранжирования кадров видеопоследовательности.

Для анализа вклада отдельных компонент рассмотрим декомпозицию итогового значения кри-

терия $S(I)$. В данном примере весовые коэффициенты приняты равными единице ($\alpha_l = 1, \beta_l = 1$), что позволяет непосредственно оценить относительный вклад энергетических и межуровневых компонент критерия. Результаты декомпозиции представлены в табл. 1.

Табл. 1. Вклад отдельных компонент в итоговую оценку структурной информативности $S(I)$

Компонент	Значение	Вклад, %
$\tilde{E}_1$	0.013690	0.7
$\tilde{E}_2$	0.169649	8.8
$\tilde{E}_3$	1.275887	66.5
C_1	0.183256	9.6
C_2	0.275206	14.4
$S(I)$	1.917688	100

Из данных, приведенных в табл. 1, следует, что основной вклад в значение критерия $S(I)$ вносит энергетическая компонента третьего уровня разложения $\tilde{E}_3$, доля которой составляет около 66.6% от общего значения оценки. Это указывает на то, что значительная часть структурной информации сцены представлена на более крупных пространственных масштабах. Подобное распределение энергии характерно для изображений городской инфраструктуры. В рассматриваемой сцене присутствуют выраженные геометрические структуры — дорожные линии, перекрестки и границы городских кварталов, — которые проявляются преимущественно на более грубых уровнях вейвлет-разложения.

Суммарный вклад компонент согласованности между уровнями разложения C_l составляет около 24% от итогового значения критерия, что свидетельствует о наличии пространственных структур, устойчиво проявляющихся на соседних уровнях разложения.

При необходимости может быть введено дополнительное ограничение на минимальный временной интервал между выбранными кадрами для обеспечения разнообразия ракурсов наблюдения. Стоит отметить, что вычисление оценки $S(I)$ не требует информации о спутниковом изображении и может быть выполнено на ранних этапах обработки.

Вычислительная сложность

Предлагаемый метод имеет вычислительную сложность $O(HW)$ на кадр, где H и W обозначают соответственно высоту и ширину изображения в пикселях. Таким образом, величина HW соответствует общему числу пикселей изображения. При стандартной пирамидальной реализации 2D-вейвлет-разложения и фиксированной длине фильтра сложность алгоритма остается линейной по числу пикселей.

Линейность обеспечивается тем, что суммарный объем вычислений по всем уровням разложения убывает геометрически, поскольку каждый последующий уровень применяется к аппроксимирующему поддиапазону, имеющему вдвое меньший размер по каждой координате. В результате число операций на уровне l пропорционально $HW / 4^l$, а суммарная сложность по всем уровням остается порядка $O(HW)$.

Таким образом, при обработке видеопоследовательности из T кадров общая вычислительная сложность этапа структурного анализа составляет $O(T \cdot HW)$.

По сравнению с последующим этапом кросс-модального сопоставления изображений вычислительные затраты на вейвлет-анализ являются незначительными. В частности, для современных моделей плотного сопоставления [17, 18], основанных на построении глобальной корреляционной матрицы или механизме внимания, вычислительная сложность может достигать $O(N^2)$, по числу токенов или пространственных позиций. Здесь N обозначает количество элементов пространственного представления изображения (признаковых точек после дискретизации признакового пространства). Поскольку N обычно пропорционально числу пространственных позиций признаковой карты, квадратичная зависимость приводит к значительно более высоким вычислительным затратам по сравнению с линейной сложностью предварительного вейвлет-анализа.

ОПИСАНИЕ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Видеоизображения

Основной целью экспериментов является анализ влияния предварительного отбора кадров на точность сопоставления, устойчивость регистрации и вычислительную эффективность системы в целом. В экспериментальном исследовании использовались открытые наборы данных, содержащие изображения, полученные с беспилотных летательных аппаратов, а также спутниковые снимки соответствующих участков местности. Основная часть экспериментов проводилась на данных датасета UAV-VisLoc [19], предназначенного для задачи визуальной локализации БПЛА по спутниковым изображениям. Данный датасет содержит пары изображений: полученные с борта БПЛА и соответствующие спутниковые снимки одной и той же территории. Изображения характеризуются значительными вариациями масштаба, ракурса наблюдения и структурного состава сцены, что делает задачу сопоставления существенно более сложной. Для анализа поведения методов в инфракрасном диапазоне дополнительно использовались изображения из датасета VisDrone-DroneVehicle [20]. Данный набор данных содержит изображения, полученные с беспилотных летательных аппаратов в видимом и инфракрасном диапазонах спектра. Поскольку инфракрасные изображения данного датасета не сопровождаются точной географической привязкой, соответствующие спутниковые фрагменты для них подбирались эмпирически на основе визуального сходства структуры сцены. С целью моделирования



Рис. 3. Примеры кадров экспериментальной выборки.
а, б — кадры в видимом диапазоне; в, г — кадры в инфракрасном диапазоне

условий сложного сценария полета БПЛА, связанных с резкими маневрами и ухудшением качества изображения, к части кадров применялись методы искусственной деградации изображения. В частности, использовались операции размытия, добавления засветки и аддитивного шума, имитирующие эффекты движения камеры, изменения освещенности и деградации сенсора. Такие преобразования позволяют приблизить экспериментальные условия к реальным полетным сценариям, при которых наблюдаются резкие изменения поля зрения камеры, геометрические искажения и снижение контрастности изображения. Сформированная экспериментальная выборка содержит кадры с существенно различной структурной информативностью. Часть изображений содержит выраженные геометрические элементы сцены, такие как границы объектов, дорожная сеть и контрастные текстуры, тогда как другие характеризуются слабой структурной выраженностью или значительными искажениями. Это позволяет

объективно оценить способность предлагаемого метода предварительного отбора выявлять кадры, обладающие наибольшим потенциалом для успешного кросс-модального сопоставления со спутниковыми изображениями. Примеры кадров экспериментальной выборки представлены на рис. 3.

Протокол экспериментальной оценки

Для обеспечения корректности сравнения все методы предварительного отбора кадров оценивались в идентичном вычислительном протоколе. Эксперименты проводились с использованием двух моделей кросс-модального сопоставления: LoFTR [17] и RoMa [18]. Поскольку используемые открытые датасеты не содержат видеопоследовательностей в явном виде, для проведения экспериментов из изображений датасетов UAV-VisLoc и VisDrone-DroneVehicle были сформированы искусственные видеопоследовательности.

Оценка выполнялась на десяти видеопоследовательностях, каждая из которых содержала

60 кадров. Последовательности 1–7 соответствовали изображениям, полученным в видимом диапазоне спектра и сформированным на основе данных датасета UAV-VisLoc. Последовательности 8–10 содержали инфракрасные изображения и сформированы на основе датасета VisDrone-DroneVehicle.

Для каждой модели сопоставления все стратегии предварительного отбора кадров сравнивались при неизменной архитектуре сети, фиксированных параметрах и одинаковых весах модели. Это позволяло исключить влияние параметров сопоставления и сосредоточиться на анализе эффективности методов предварительного отбора.

Геометрическая верификация выполнялась с использованием алгоритма RANSAC [21] с фиксированными параметрами, включая максимальное число итераций и порог репроекционной ошибки τ . В результате различия в числе найденных инлайеров* определяются исключительно качеством предварительного отбора кадров при прочих равных условиях сопоставления и геометрической фильтрации.

Метрики оценки

Типы метрик

Пусть последовательность содержит N кадров:

$$\{(S_i, M_i)\}_{i=1}^N,$$

где S_i — значение структурного показателя для i -го кадра, а M_i — число инлайеров, полученных после этапа геометрической верификации.

Для количественной оценки эффективности предложенного подхода используются две группы метрик:

1) корреляционные метрики, характеризующие статистическую зависимость между S_i и M_i ;

2) метрики эффективности отбора, отражающие практический выигрыш при выборе топ K кадров.

Корреляционные метрики

Данная группа метрик предназначена для анализа статистической зависимости между значением структурного показателя S_i и числом инлайеров M_i . В работе используются два коэффициента корреляции.

Линейная зависимость оценивается коэффициентом корреляции Пирсона:

$$r = \frac{\text{cov}(S_i, M_i)}{\sigma_S \sigma_M},$$

где $\text{cov}(\cdot)$ — ковариация, σ_S и σ_M — стандартные отклонения величин S_i и M_i соответственно.

Монотонная зависимость оценивается коэффициентом ранговой корреляции Спирмена:

$$\rho = \text{corr}_{\text{rank}}(S_i, M_i).$$

Для каждого коэффициента вычисляется соответствующее значение p , характеризующее статистическую значимость выявленной зависимости.

В рамках проверки статистической гипотезы рассматривается нулевая гипотеза H_0 об отсутствии корреляционной связи между величинами S_i и M_i . Значение p определяется как вероятность получить наблюдаемое значение коэффициента корреляции при условии справедливости H_0 .

При $p < \alpha$ нулевая гипотеза отклоняется и выявленная зависимость считается статистически значимой. В работе используется уровень значимости $\alpha = 0.05$.

Метрики эффективности отбора кадров

Данная группа метрик предназначена для оценки практической полезности предложенного критерия при выборе подмножества кадров.

Отбор Топ- K кадров. Обозначим через T_k множество из k кадров с наибольшими значениями показателя S_i :

$$T_k = \text{Top-}K(S).$$

Среднее число инлайеров для выбранных кадров определяется как:

$$\mu_{\text{top}} = \frac{1}{k} \sum_{i \in T_k} M_i.$$

В качестве базового уровня рассматривается случайный выбор k кадров из последовательности. Пусть R_k — случайное подмножество мощности k . Оценка среднего числа инлайеров при случайном выборе вычисляется методом Монте-Карло:

$$\mu_{\text{rand}} = E_{R_k} \left[\frac{1}{k} \sum_{i \in R_k} M_i \right],$$

в дальнейшем $Random = \mu_{\text{rand}}$.

* Под инлайерами понимаются геометрически согласованные соответствия, удовлетворяющие выбранной модели преобразования.

Абсолютный выигрыш от использования предложенного критерия определяется как:

$$\Delta @ k = \mu_{\text{top}} - \mu_{\text{rand}}.$$

Относительный выигрыш по сравнению со случайным выбором определяется как:

$$\text{Ratio} = \frac{\mu_{\text{top}}}{\mu_{\text{rand}}}.$$

ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

Табличное представление результатов

В данном разделе приводится анализ результатов оценки предложенного критерия предварительного отбора кадров. Поскольку абсолютное число инлайеров M_i существенно зависит от используемой модели сопоставления, эксперименты проводились для двух архитектур матчинга: ELoFTR и RoMa. В экспериментах использовались

предобученные веса моделей из работы MatchAnything [22]. Эти модели представляют собой архитектуры ELoFTR и RoMa, дополнительно адаптированные для задачи кросс-модального сопоставления изображений (БПЛА $\leftrightarrow$ спутник). В процессе обучения параметры сетей были оптимизированы с учетом различий в спектральных характеристиках, контрасте и масштабе между аэровизуальными и спутниковыми данными. Таким образом, сравнение проводится не между исходными универсальными моделями, а между их специализированными версиями, адаптированными для условий кросс-модального сопоставления.

Эксперименты проводились на десяти видеопоследовательностях (Sequence), каждая из которых содержала $N = 60$ кадров. Из каждой последовательности отбиралось $k = 15$ наиболее информативных кадров согласно предложенному структурному критерию. Сводные количественные результаты экспериментов представлены в табл. 2, 3.

Табл. 2. Результаты оценки качества отбора кадров для модели на основе архитектуры ELoFTR

Sequence	N	k	r	ρ	Top- k	Random	$\Delta @ k$	Ratio
1	60	15	0.966	0.760	2095.3	1286.0	809.3	1.63
2	60	15	0.980	0.816	3445.7	2248.5	1197.2	1.53
3	60	15	0.420	0.548	932.3	324.5	607.8	2.87
4	60	15	0.992	0.721	3446.3	2403.9	1042.4	1.43
5	60	15	0.979	0.515	3160.7	2351.5	809.2	1.34
6	60	15	0.990	0.418	3501.0	2588.5	912.5	1.35
7	60	15	0.989	0.697	3711.0	2521.3	1189.7	1.47
8	60	15	0.950	0.418	981.3	764.8	216.5	1.28
9	60	15	0.960	0.491	983.4	746.4	237.0	1.32
10	60	15	0.973	0.406	959.6	732.8	226.8	1.31

Табл. 3. Результаты оценки качества отбора кадров для модели на основе архитектуры RoMa

Sequence	N	k	r	ρ	Top- k	Random	$\Delta @ k$	Ratio
1	60	15	0.578	0.552	8353.3	8045.5	307.9	1.04
2	60	15	0.965	0.564	8008.3	6954.7	1053.6	1.15
3	60	15	0.723	0.909	8326.7	4804.6	3522.1	1.73
4	60	15	0.883	0.879	8201.3	7703.9	497.4	1.06
5	60	15	0.065	0.164	7956.3	7860.3	96.0	1.01
6	60	15	0.729	0.479	7851.7	7728.3	123.4	1.02
7	60	15	0.966	0.491	8274.7	7264.8	1009.9	1.14
8	60	15	0.586	0.505	2552.2	2546.1	6.1	1.00
9	60	15	0.892	0.515	2518.5	2324.1	194.4	1.08
10	60	15	0.824	0.709	2716.3	2510.8	205.5	1.08

Анализ данных табл. 2

Табл. 2 содержит результаты для архитектуры ELoFTR, тогда как табл. 3 соответствует архитектуре RoMa. В таблицах для каждой последовательности (Sequence) приведены значения корреляционных метрик (r , ρ) и показателей практической эффективности отбора (Top- k , Random, $\Delta@k$, Ratio). Корреляционные метрики характеризуют степень согласованности структурного показателя с фактической успешностью сопоставления, в то время как метрики Top- k , $\Delta@k$, Ratio отражают практическую эффективность отбора кадров.

Статистическая значимость коэффициентов корреляции оценивалась по соответствующим p -значениям. Значимыми считались коэффициенты, для которых выполнялось условие $p < 0.05$.

Как следует из табл. 2, в большинстве последовательностей наблюдается высокая положительная корреляция между значением структурного показателя и числом инлайеров. Для девяти из десяти последовательностей коэффициент Пирсона превышает 0.95, что свидетельствует о сильной линейной зависимости между оценкой структурной информативности кадра и успешностью сопоставления. Исключением является последовательность 3 ($r = 0.420$, $\rho = 0.548$), где корреляционная связь выражена заметно слабее.

Высокие значения корреляционных метрик указывают на то, что предложенный структурный показатель адекватно отражает вероятность успешного сопоставления кадра со спутниковым изображением. Иными словами, кадры с более выраженной структурной организацией на разных масштабах, как правило, обеспечивают большее число устойчивых геометрических соответствий после этапа верификации. Это подтверждает гипотезу о том, что структурная информативность изображения является важным показателем успешности кросс-модального сопоставления.

Практическая эффективность предложенного критерия отбора подтверждается значениями показателя Ratio, который для всех последовательностей превышает единицу. Это означает, что отбор кадров по структурному критерию стабильно превосходит случайный выбор. Относительный выигрыш варьируется в диапазоне от 1.28 до 2.87. Наиболее выраженный эффект наблюдается в последовательности 3 (Ratio 2.87), несмотря на относительно низкие значения корреляционных метрик. Данный результат указывает на то, что даже при ослаблении глобальной корреляционной зависимости критерий способен выделять небольшое подмножество кадров, содержащих наиболее информативные структурные элементы сцены. Вероятной

причиной такого поведения является неоднородность структуры видеопоследовательности, когда лишь небольшая часть кадров содержит геометрически устойчивые элементы (дороги, границы застройки, контуры объектов), обеспечивающие высокое число соответствий.

Даже в последовательностях с умеренными значениями ранговой корреляции (например, последовательности 5–10) практический выигрыш остается заметным. Это свидетельствует о том, что предложенный критерий выполняет не только функцию ранжирования кадров, но и эффективно концентрирует вычислительные ресурсы на наиболее информативных фрагментах видеопоследовательности, что особенно важно в условиях ограниченного вычислительного бюджета бортовых систем.

Анализ данных табл. 3

В табл. 3 приведены аналогичные результаты для модели RoMa. Абсолютные значения числа инлайеров в данной конфигурации существенно выше по сравнению с архитектурой ELoFTR, что отражает более высокую плотность найденных соответствий. Это согласуется с особенностями архитектуры RoMa, ориентированной на получение более плотных и пространственно распределенных соответствий.

Для большинства последовательностей наблюдается положительная корреляция между структурным показателем и числом инлайеров, однако степень этой зависимости варьируется заметно сильнее, чем в случае ELoFTR. Наиболее высокая ранговая корреляция наблюдается в последовательностях 3 и 4 ($\rho \approx$ от 0.88 до 0.91), тогда как в отдельных последовательностях (например, последовательность 5) корреляционная связь практически отсутствует. Это может быть связано с тем, что RoMa менее чувствительна к локальным структурным особенностям изображений и способна находить соответствия даже в условиях слабо выраженной геометрической структуры.

Практический выигрыш предварительного отбора для RoMa в среднем является более умеренным. Значения Ratio в большинстве последовательностей находятся в диапазоне от 1.01 до 1.15, что указывает на сравнительно небольшое, но стабильное превосходство структурного отбора над случайным выбором. Наиболее выраженный эффект наблюдается в последовательности 3 (Ratio 1.73), где структурный критерий позволяет существенно увеличить число найденных инлайеров. В последовательностях 5 и 8 выигрыш практически отсутствует, что может быть связано со структурной однородностью сцены или малой вариативностью кадров внутри последовательности.

Сравнение двух конфигураций позволяет сделать несколько важных наблюдений. Во-первых, для обеих моделей предложенный структурный критерий демонстрирует положительную связь с успешностью сопоставления, что подтверждает его модельно-независимый характер и переносимость между различными архитектурами матчинга. Во-вторых, для конфигурации ELoFTR наблюдается более выраженный выигрыш от предварительного отбора кадров. Это указывает на более высокую чувствительность данной архитектуры к качеству входного набора изображений и подчеркивает важность этапа предварительной фильтрации при использовании менее плотных методов сопоставления.

Для RoMa влияние предварительного отбора оказывается менее контрастным, что может быть связано с большей устойчивостью архитектуры к вариациям структурной информативности кадров. Благодаря более плотному механизму сопоставления RoMa способна извлекать соответствия даже из менее информативных изображений, вследствие чего эффект предварительного отбора проявляется слабее. Тем не менее положительный выигрыш сохраняется в большинстве последовательностей, что подтверждает практическую целесообразность использования структурного критерия в качестве этапа предварительной фильтрации кадров видеопотока.

Сравнение с базовыми метриками качества изображения

Для оценки преимуществ предложенного критерия предварительного отбора было выполнено его сравнение с рядом широко используемых метрик информативности изображения. В качестве базовых методов рассматривались: энтропия, градиентная метрика, метрика лапласиана и метрика тензора структуры. Выбор данных показателей обусловлен тем, что они отражают различные аспекты визуальной информативности кадра: энтропия характеризует общее разнообразие яркостных значений, градиент и лапласиан — выраженность локальных перепадов яркости и контурных структур, а тензор структуры — интенсивность и направленную согласованность локальных градиентов.

Каждая из перечисленных метрик использовалась в качестве критерия ранжирования кадров видеопоследовательности аналогично предложенному вейвлет-критерию. Иными словами, для каждой последовательности кадры сортировались по убыванию значения соответствующего показателя, после чего выбиралось подмножество из k лучших кадров, на котором далее оценивалась успешность сопоставления со спутниковым изображением.

Во всех экспериментах для сравнительных методов использовались одинаковые параметры отбора: объем последовательности составлял $N = 60$ кадров, а число отбираемых кадров $k = 15$. Поскольку значение *Random* является общим для всех методов внутри одной последовательности, в сравнительных таблицах оно не приводится. Для каждого метода указывались коэффициенты корреляции Пирсона r и Спирмена ρ , а также показатели практической эффективности отбора: $\text{Top-}k$, $\Delta@k$ и *Ratio*.

Сравнение проводилось отдельно для двух архитектур матчинга: ELoFTR и RoMa. Такое раздельное рассмотрение необходимо, поскольку эффективность предварительного отбора может зависеть от особенностей используемой модели сопоставления. Результаты сравнения представлены в Приложении в табл. П1 и П2. Табл. П1 содержит результаты для архитектуры ELoFTR, а табл. П2 — результаты для архитектуры RoMa. Такое представление позволяет оценить не только абсолютное качество предложенного критерия, но и его устойчивость относительно различных схем кросс-модального сопоставления.

Как следует из табл. П1, для архитектуры ELoFTR предложенный вейвлет-критерий демонстрирует наиболее высокую согласованность с числом инлайеров в большинстве рассмотренных последовательностей. В частности, в последовательностях 01, 02, 04–08 он обеспечивает максимальные или близкие к максимальным значения коэффициента Пирсона r , что указывает на его высокую способность отражать изменение успешности сопоставления при переходе от кадра к кадру. Иными словами, вейвлет-представление лучше других рассматриваемых метрик описывает количественную связь между структурной выраженностью изображения и числом корректных геометрических соответствий.

При этом по практическим показателям $\text{Top-}k$, $\Delta@k$ и *Ratio* преимущество вейвлет-критерия носит более неоднородный характер. В ряде последовательностей простые градиентные метрики (градиент, лапласиан, тензор структуры) демонстрируют сопоставимые или даже более высокие значения практического выигрыша. Это особенно заметно в последовательностях 01, 06 и 08, где локальные меры контрастности и контурной выраженности оказываются достаточно информативными для отбора кадров. Тем не менее вейвлет-критерий показывает существенное преимущество в последовательностях 02, 03 и 04, где он обеспечивает наибольшие значения $\text{Top-}k$, $\Delta@k$ и *Ratio*. Наиболее показательной является последовательность 03: если для всех базовых методов значение *Ratio* составляет около 1.67, то для вейвлет-критерия оно возрастает до 2.87. Это указывает

на то, что в условиях структурно неоднородной сцены многошкальный анализ позволяет выделять действительно информативные кадры, тогда как простые локальные показатели оказываются недостаточно селективными.

Следует также отметить, что для ряда последовательностей градиент, лапласиан и тензор структуры дают практически одинаковые значения $\text{Top-}k$, $\Delta@k$ и Ratio . Такое поведение закономерно, поскольку все три показателя основаны на анализе локальных градиентов и в значительной степени отражают один и тот же тип структурной информации, а именно интенсивность границ и контуров. В отличие от них, предложенный вейвлет-критерий учитывает распределение структурной энергии по нескольким масштабам, что обеспечивает более надежное различение кадров в сложных сценах.

Результаты для архитектуры RoMa, приведенные в табл. П2, демонстрируют иную картину. С одной стороны, вейвлет-критерий и здесь показывает высокие значения корреляционных метрик в ряде последовательностей, особенно в последовательностях 03, 04 и 07. С другой стороны, его преимущество по практическим показателям становится менее устойчивым. Во многих последовательностях градиентные метрики обеспечивают сопоставимый или более высокий выигрыш, а различия между методами оказываются существенно меньше, чем в случае ELoFTR.

Наиболее яркий результат для RoMa наблюдается в последовательности 03. Если энтропия, градиент, лапласиан и тензор структуры приводят к снижению качества отбора относительно случайного выбора ($\text{Ratio} \approx 0.82$), то вейвлет-критерий, напротив, обеспечивает выраженное улучшение ($\text{Ratio} 1.73$, $\rho = 0.99$). Это свидетельствует о том, что в сложных последовательностях подход на основе вейвлет способен выявлять устойчивые многошкальные структуры, которые оказываются значимыми и для более мощной архитектуры сопоставления. Однако в большинстве остальных последовательностей выигрыш вейвлет-критерия для RoMa либо невелик, либо уступает более простым базовым метрикам.

Сравнение двух таблиц позволяет сделать несколько обобщающих выводов. Во-первых, предложенный вейвлет-критерий обладает модельно-независимой полезностью: для обеих архитектур он демонстрирует положительную связь с успешностью сопоставления и в отдельных последовательностях обеспечивает наилучший практический результат. Во-вторых, его преимущество наиболее ярко проявляется для архитектуры ELoFTR, то есть для модели, более чувствительной к качеству предварительно отобранного набора кадров. В этом случае анализ структуры изо-

бражения на разных масштабах действительно дает преимущество по сравнению с простыми локальными метриками. В-третьих, для RoMa влияние предварительного отбора оказывается менее выраженным, что, вероятно, связано с большей внутренней устойчивостью данной архитектуры к вариациям структурной информативности кадров.

Таким образом, предложенный критерий целесообразно рассматривать не просто как альтернативу традиционным метрикам качества изображения, а как специализированный разномасштабный показатель, ориентированный именно на задачу предварительного отбора кадров для последующего кросс-модального сопоставления. Его основное преимущество проявляется в сложных сценах, где локальные контрастные характеристики не позволяют надежно отделить действительно информативные кадры от менее полезных.

ЗАКЛЮЧЕНИЕ

В работе рассмотрена задача предварительного отбора информативных кадров для систем кросс-модального сопоставления изображений БПЛА и спутника в условиях ограниченных вычислительных ресурсов. Высокая частота формирования видеопотока в системах визуальной навигации БПЛА приводит к значительному объему входных данных, тогда как вычислительная сложность современных алгоритмов сопоставления ограничивает число кадров, которые могут быть обработаны в режиме реального времени. В этих условиях возникает необходимость приоритизации входных изображений и выделения подмножества кадров, обладающих наибольшей пригодностью для последующего сопоставления.

Предложен структурный критерий отбора кадров, основанный на анализе структуры изображения на разных масштабах и согласованности ориентированных компонент. В отличие от традиционных метрик качества изображения общего назначения, предложенный показатель ориентирован на выявление геометрически устойчивых структур, которые играют ключевую роль при установлении соответствий между изображениями.

Экспериментальная оценка проводилась на последовательностях, сформированных из открытых датасетов UAV-VisLoc и VisDrone-DroneVehicle для двух конфигураций кросс-модального сопоставления: ELoFTR и RoMa. Полученные результаты демонстрируют устойчивую положительную корреляцию между значением структурного показателя и числом инлайеров, что подтверждает связь разномасштабной структурной информативности кадра с успешностью процедуры сопоставления.

Показано, что предварительный отбор кадров на основе предложенного критерия обеспечивает увеличение числа корректных соответствий по сравнению со случайным выбором кадров. Для конфигурации ELoFTR выигрыш достигает от 1.3 до 1.8 раза, что свидетельствует о высокой чувствительности данной архитектуры к качеству входного набора изображений. Для модели RoMa эффект отбора выражен слабее, однако в большинстве последовательностей также наблюдается положительный прирост числа найденных соответствий.

Сравнение с рядом базовых метрик информативности изображения показало, что предложенный вейвлет-критерий обеспечивает более устойчивую корреляцию с успешностью сопоставления и демонстрирует преимущество в сценах со сложной структурной организацией. Это подтверждает

целесообразность использования разномасштабного анализа для оценки пригодности кадров в задачах визуальной навигации. Таким образом, предложенный подход может рассматриваться как вычислительно эффективный механизм предварительной выборки кадров в системах визуальной навигации БПЛА. Его применение позволяет повысить устойчивость кросс-модального сопоставления без увеличения вычислительных затрат на этапе матчинга.

В дальнейшем представляется перспективным исследование адаптивных стратегий отбора кадров, учитывающих динамику сцены и параметры движения, а также интеграция предложенного структурного показателя непосредственно в процесс обучения кросс-модальных моделей сопоставления.

ПРИЛОЖЕНИЕ

Табл. П1. Сравнение предложенного вейвлет-критерия с популярными метриками качества изображения для архитектуры ELoFTR (для каждой последовательности приведены корреляционные метрики и показатели практической эффективности отбора кадров)

Sequence	Метод	r	ρ	Top- k	$\Delta@k$	Ratio
01	Энтропия	0.884	0.815	2148.3	862.3	1.67
	Градиент	0.756	0.851	2223.3	937.3	1.73
	Лапласиан	0.870	0.881	2223.3	937.3	1.73
	Тензор структуры	0.836	0.863	2223.3	937.3	1.73
	Вейвлет	0.966	0.760	2095.3	809.3	1.63
02	Энтропия	0.918	0.558	3108.3	859.8	1.38
	Градиент	0.957	0.791	3265.3	1016.8	1.45
	Лапласиан	0.953	0.791	3265.3	1016.8	1.45
	Тензор структуры	0.950	0.791	3265.3	1016.8	1.45
	Вейвлет	0.980	0.816	3445.7	1197.2	1.53
03	Энтропия	-0.008	0.333	540.7	216.1	1.67
	Градиент	0.206	0.569	540.7	216.1	1.67
	Лапласиан	0.245	0.569	540.7	216.1	1.67
	Тензор структуры	0.224	0.569	540.7	216.1	1.67
	Вейвлет	0.420	0.548	932.3	607.8	2.87
04	Энтропия	0.919	0.491	3382.7	978.8	1.41
	Градиент	0.710	0.455	3382.7	978.8	1.41
	Лапласиан	0.730	0.479	3382.7	978.8	1.41
	Тензор структуры	0.673	0.406	3382.7	978.8	1.41
	Вейвлет	0.992	0.721	3446.3	1042.4	1.43
05	Энтропия	0.933	0.733	3278.7	927.2	1.39
	Градиент	0.849	0.600	3125.7	774.2	1.33
	Лапласиан	0.878	0.600	3125.7	774.2	1.33
	Тензор структуры	0.761	0.576	3125.7	774.2	1.33
	Вейвлет	0.979	0.515	3160.7	809.2	1.34

Табл. П1. (окончание)

Sequence	Метод	r	ρ	Top- k	$\Delta@k$	Ratio
06	Энтропия	-0.121	0.285	3419.3	830.9	1.32
	Градиент	0.658	0.818	3419.3	830.9	1.32
	Лапласиан	0.718	0.745	3419.3	830.9	1.32
	Тензор структуры	0.634	0.915	3547.0	958.5	1.37
	Вейвлет	0.990	0.418	3501.0	912.5	1.35
07	Энтропия	0.823	0.600	3506.0	984.7	1.39
	Градиент	0.911	0.758	3711.0	1189.7	1.47
	Лапласиан	0.910	0.661	3711.0	1189.7	1.47
	Тензор структуры	0.889	0.733	3711.0	1189.7	1.47
	Вейвлет	0.989	0.697	3711.0	1189.7	1.47
08	Энтропия	0.428	0.103	1022.8	257.9	1.34
	Градиент	0.802	0.648	1044.0	279.1	1.36
	Лапласиан	0.813	0.636	1022.8	257.9	1.34
	Тензор структуры	0.771	0.552	1044.0	279.1	1.36
	Вейвлет	0.950	0.418	981.3	216.5	1.28

Табл. П2. Сравнение предложенного вейвлет-критерия с популярными метриками качества изображения для архитектуры RoMa (для каждой последовательности приведены корреляционные метрики и показатели практической эффективности отбора кадров)

Sequence	Метод	r	ρ	Top- k	$\Delta@k$	Ratio
01	Энтропия	0.523	0.430	8342.3	296.9	1.04
	Градиент	0.675	0.770	8901.3	855.9	1.11
	Лапласиан	0.586	0.661	8901.3	855.9	1.11
	Тензор структуры	0.490	0.624	8901.3	855.9	1.11
	Вейвлет	0.578	0.552	8353.3	307.9	1.04
02	Энтропия	0.903	0.697	8197.0	1242.3	1.18
	Градиент	0.956	0.648	8323.3	1368.6	1.20
	Лапласиан	0.945	0.648	8323.3	1368.6	1.20
	Тензор структуры	0.947	0.697	8323.3	1368.6	1.20
	Вейвлет	0.965	0.564	8008.3	1053.6	1.15
03	Энтропия	-0.169	-0.108	3958.7	-845.9	0.82
	Градиент	0.275	0.439	3958.7	-845.9	0.82
	Лапласиан	0.320	0.439	3958.7	-845.9	0.82
	Тензор структуры	0.258	0.439	3958.7	-845.9	0.82
	Вейвлет	0.723	0.909	8326.7	3522.1	1.73
04	Энтропия	0.745	0.770	7929.0	225.1	1.03
	Градиент	0.478	0.685	7929.0	225.1	1.03
	Лапласиан	0.516	0.733	7929.0	225.1	1.03
	Тензор структуры	0.433	0.636	7929.0	225.1	1.03
	Вейвлет	0.883	0.879	8201.3	497.4	1.06

Табл. П2. (окончание)

05	Энтропия	-0.063	-0.152	7623.3	-237.0	0.97
	Градиент	0.196	0.248	8347.0	486.7	1.06
	Лапласиан	0.181	0.309	8347.0	486.7	1.06
	Тензор структуры	0.258	0.394	8347.0	486.7	1.06
	Вейвлет	0.065	0.164	7956.3	96.0	1.01
06	Энтропия	0.134	0.309	8228.0	499.7	1.06
	Градиент	0.664	0.576	8228.0	499.7	1.06
	Лапласиан	0.704	0.612	8228.0	499.7	1.06
	Тензор структуры	0.640	0.588	7995.0	266.7	1.03
	Вейвлет	0.729	0.479	7851.7	123.4	1.02
07	Энтропия	0.821	0.503	8301.0	1036.2	1.14
	Градиент	0.890	0.576	8274.7	1009.9	1.14
	Лапласиан	0.880	0.636	8274.7	1009.9	1.14
	Тензор структуры	0.865	0.564	8274.7	1009.9	1.14
	Вейвлет	0.966	0.491	8274.7	1009.9	1.14
08	Энтропия	-0.047	-0.255	2521.5	-24.6	0.99
	Градиент	0.299	0.340	2597.2	51.1	1.02
	Лапласиан	0.337	0.267	2521.5	-24.6	0.99
	Тензор структуры	0.289	0.401	2597.2	51.1	1.02
	Вейвлет	0.586	0.505	2552.2	6.1	1.00

СПИСОК ИСТОЧНИКОВ

1. Mohsan S.A.H., Othman N.Q.H., Li Y., Alsharif M.H. Unmanned aerial vehicles (UAVs): Practical aspects, applications, open challenges, security issues, and future trends // *Intelligent service robotics*. 2023. Vol. 16, no. 1. P. 109–137. DOI: 10.1007/s11370-022-00452-4
2. Teed Z., Deng J. Raft: Recurrent all-pairs field transforms for optical flow // *European conference on computer vision*. Cham: Springer International Publishing. 2020. P. 402–419. DOI: 10.48550/arXiv.2003.12039
3. Forster C., Zhang Z., Gassner M., Werlberger M., Scaramuzza D. SVO: Semidirect visual odometry for monocular and multicamera systems // *IEEE Transactions on Robotics*. 2016. Vol. 33, no. 2. P. 249–265. DOI: 10.1109/TRO.2016.2623335
4. Crete F., Dolmiere T., Ladret P., Nicolas M. The blur effect: perception and estimation with a new no-reference perceptual blur metric // *Human vision and electronic imaging XII*. SPIE. 2007. Vol. 6492. P. 196–206. DOI: 10.1117/12.702790
5. Eskicioglu A.M., Fisher P.S. Image quality measures and their performance // *IEEE Transactions on communications*. 1995. Vol. 43, no. 12. P. 2959–2965. DOI: 10.1109/26.477498
6. Гаврилов Д.А., Ловцов Д.А., Татаринова Е.А. Эффективная автоматизированная переработка визуальной информации в оптико-электронной системе наземно космического мониторинга // *Профессорский журнал. Сер. Технические науки*. 2019. № 3. С. 26–40. URL: <https://www.elibrary.ru/item.asp?id=44517909>
7. Revaud J., De Souza C., Humenberger M., Weinzaepfel P. R2d2: Reliable and repeatable detector and descriptor // *Advances in neural information processing systems*. 2019. Vol. 32. Id. 6717. URL: <https://dl.acm.org/doi/10.5555/3454287.3455400>
8. Ahmed F., Jenihhin M. A survey on UAV computing platforms: A hardware reliability perspective // *Sensors*. 2022. Vol. 22, no. 16. Id. 6286. DOI: 10.3390/s22166286
9. Лапушкин А.Г., Гаврилов Д.А., Поткин О.А. Программное обеспечение для создания синтезированных данных и симулятор с обратной связью для тестирования алгоритмов машинного обучения // *Научное приборостроение*. 2023. Т. 33, № 1. С. 95–108. URL: <http://iairas.ru/mag/2023/abst1.php#abst8>
10. Mallat S.G. A theory for multiresolution signal decomposition: the wavelet representation // *IEEE Trans. Pattern Anal. Mach. Intell.* 1989. Vol. 11, no. 7. P. 674–693. DOI: 10.1109/34.192463
11. Гаврилов Д.А., Щелкунов Н.Н. Программное обеспечение разметки крупноформатных аэрокосмических изображений и подготовки обучающих выборок // *Научное приборостроение*. 2020. Vol. 30, № 2. С. 67–75. URL: <http://iairas.ru/mag/2020/abst2.php#abst9>
12. Bracewell R.N. The Fourier transform and its applications. New York, McGraw-hill, 1986. 422 p. URL: <https://see.stanford.edu/materials/lsofaee261/book-fall-07.pdf>
13. Shannon C.E. A mathematical theory of communication // *The Bell system technical journal*. 1948. Vol. 27, no. 3. P. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x

14. *Tenenbaum J.M.* Discrimination of Visual Patterns by Gradient Methods // *Information and Control*. 1970. Vol. 17, no. 1. P. 21–43.
15. *Pertuz S., Puig D., Garcia M.A.* Analysis of focus measure operators for shape-from-focus // *Pattern Recognition*. 2013. Vol. 46, no. 5. P. 1415–432. DOI: 10.1016/j.patcog.2012.11.011
16. *Bigun J., Granlund G.H.* Optimal orientation detection of linear symmetry // *Proceedings of the IEEE First International Conference on Computer Vision*. 1987. Vol. 54. P. 433–438. URL: https://www2.hh.se/staff/josef/publ/publications/bigun86tr_london.pdf
17. *Sun J., Shen Z., Wang Y., Bao H., Zhou X.* LoFTR: Detector-free local feature matching with transformers // *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021. P. 8922–8931. DOI: 10.48550/arXiv.2104.00680
18. *Edstedt J., Sun Q., Bökman G., Wadenbäck M., Felsberg M.* Roma: Robust dense feature matching // *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024. P. 19790–19800. DOI: 10.48550/arXiv.2305.15404
19. *Xu W., Yao Y., Cao J., Wei Z., Liu C., Wang J., Peng M.* Uav-visloc: A large-scale dataset for uav visual localization // *arXiv:2405.11936 [cs.CV]*. 2024. DOI: 10.48550/arXiv.2405.11936
20. *Tang P., Li J., Sun H.* A Review of Electric UAV Visual Detection and Navigation Technologies for Emergency Rescue Missions // *Sustainability*. 2024. Vol. 16, no. 5. Id. 2105. DOI: 10.3390/su16052105
21. *Fischler M.A., Bolles R.C.* Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography // *Commun. ACM*. 1981. Vol. 24, no. 6. P. 381–395. DOI: 10.1145/358669.358692
22. *He X., Yu H., Peng S., Tan D., Shen Z., Bao H., Zhou X.* Matchanything: Universal cross-modality image matching with large-scale pre-training // *arXiv:2501.07556 [cs.CV]*. 2025. DOI: 10.48550/arXiv.2501.07556

**Московский физико-технический институт
(национальный исследовательский университет),
Москва**

Контакты: *Фатеев Алексей Сергеевич*,
fateev.as@mipt.ru

Материал поступил в редакцию 19.03.2026

STRUCTURE-BASED PRESELECTION OF UNMANNED AERIAL VEHICLE VIDEO FRAMES FOR CROSS-MODAL MATCHING WITH SATELLITE IMAGERY

A. S. Fateev, D. P. Krichevets, E. L. Kiselev, D. A. Gavrilov,
A. A. Fortunatov, E. A. Tatarinova

Moscow Institute of Physics and Technology (National Research University), Dolgoprudny, Moscow, Russia

The problem of preliminary selection of the most informative video frames for the visual navigation system of unmanned aerial vehicles based on cross-modal comparison with satellite images in conditions of limited computing resources is considered. The high frequency of the video stream generation leads to a significant amount of input data, whereas the matching models have limited bandwidth. As a result, only a fraction of the frames are processed in real time, which increases the risk of missing the most informative images. It has been shown that frames containing geometrically stable structures of the scene, observed at different image scales and having a consistent orientation of the boundaries, possess a higher probability of successful comparison with satellite images. The necessity of using structural criteria oriented to the tasks of comparison is substantiated.

Keywords: unmanned aerial vehicles, cross-modal image matching, frame preselection, structural information content, wavelet transform, multiscale analysis, satellite imagery, visual navigation

REFERENCES

1. Mohsan S.A.H., Othman N.Q.H., Li Y., Alsharif M.H. Unmanned aerial vehicles (UAVs): Practical aspects, applications, open challenges, security issues, and future trends. *Intelligent service robotics*, 2023, vol. 16, no. 1, pp. 109–137. DOI: 10.1007/s11370-022-00452-4
2. Teed Z., Deng J. Raft: Recurrent all-pairs field transforms for optical flow. *European conference on computer vision*. Cham: Springer International Publishing, 2020, pp. 402–419. DOI: 10.48550/arXiv.2003.12039
3. Forster C., Zhang Z., Gassner M., Werlberger M., Scaramuzza D. SVO: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics*, 2016, vol. 33, no. 2, pp. 249–265. DOI: 10.1109/TRO.2016.2623335
4. Crete F., Dolmire T., Ladret P., Nicolas M. The blur effect: perception and estimation with a new no-reference perceptual blur metric. *Human vision and electronic imaging XII. SPIE*, 2007, vol. 6492, pp. 196–206. DOI: 10.1117/12.702790
5. Eskicioglu A.M., Fisher P.S. Image quality measures and their performance. *IEEE Transactions on communications*, 1995, vol. 43, no. 12, pp. 2959–2965. DOI: 10.1109/26.477498
6. Gavrilov D.A., Lovtsov D.A., Tatarinova E.A. [Theoretical foundations of effective processing visual information in the automated optical-electronic system of ground-space monitoring]. *Professorskii zhurnal. Seriya: tekhnicheskie nauki* [Professor's journal. Series: technical sciences], 2019, no. 3, pp. 26–40. (In Russ.). URL: <https://www.elibrary.ru/item.asp?id=44517909>
7. Revaud J., De Souza C., Humenberger M., Weinzaepfel P. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 2019, vol. 32, Id. 6717. URL: <https://dl.acm.org/doi/10.5555/3454287.3455400>
8. Ahmed F., Jenihhin M. A survey on UAV computing platforms: A hardware reliability perspective. *Sensors*, 2022, vol. 22, no. 16, Id. 6286. DOI: 10.3390/s22166286
9. Lapushkin A.G., Gavrilov D.A., Potkin O.A. [Synthesized data creation software and feedback simulator for testing machine learning algorithms]. *Nauchnoe Priborostroenie [Scientific Instrumentation]*, 2023, vol. 33, no. 1, pp. 95–108. URL: <http://iairas.ru/mag/2023/abst1.php#abst8> (In Russ.).
10. Mallat S.G. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1989, vol. 11, no. 7, pp. 674–693. DOI: 10.1109/34.192463
11. Gavrilov D.A., Schelkunov N.N. [Software for large format aerospace image marking and training samples preparation]. *Nauchnoe Priborostroenie [Scientific Instrumentation]*, 2020, vol. 30, no. 2, pp. 67–75. DOI: 10.18358/np-30-2-i6775 (In Russ.).
12. Bracewell R.N. *The Fourier transform and its applications*. New York, McGraw-hill, 1986. 422 p. URL: <https://see.stanford.edu/materials/lsoftaee261/book-fall-07.pdf>
13. Shannon C.E. A mathematical theory of communication. *The Bell system technical journal*, 1948, vol. 27, no. 3, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x
14. Tenenbaum J.M. Discrimination of Visual Patterns by Gradient Methods. *Information and Control*, 1970, vol. 17, no. 1, pp. 21–43.
15. Pertuz S., Puig D., Garcia M.A. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 2013, vol. 46, no. 5, pp. 1415–432. DOI: 10.1016/j.patcog.2012.11.011

16. Bigun J., Granlund G.H. Optimal orientation detection of linear symmetry. *Proceedings of the IEEE First International Conference on Computer Vision*, 1987, vol. 54, pp. 433–438. URL: https://www2.hh.se/staff/josef/publ/publications/bigun86tr_london.pdf
17. Sun J., Shen Z., Wang Y., Bao H., Zhou X. LoFTR: Detector-free local feature matching with transformers. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931. DOI: 10.48550/arXiv.2104.00680
18. Edstedt J., Sun Q., Bökman G., Wadenbäck M., Felsberg M. Roma: Robust dense feature matching. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 19790–19800. DOI: 10.48550/arXiv.2305.15404
19. Xu W., Yao Y., Cao J., Wei Z., Liu C., Wang J., Peng M. Uav-visloc: A large-scale dataset for UAV visual localization. *arXiv:2405.11936 [cs.CV]*, 2024. DOI: 10.48550/arXiv.2405.11936
20. Tang P., Li J., Sun H. A Review of Electric UAV Visual Detection and Navigation Technologies for Emergency Rescue Missions. *Sustainability*, 2024, vol. 16, no. 5, Id. 2105. DOI: 10.3390/su16052105
21. Fischler M.A., Bolles R.C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 1981, vol. 24, no. 6, pp. 381–395. DOI: 10.1145/358669.358692
22. He X., Yu H., Peng S., Tan D., Shen Z., Bao H., Zhou X. Matchanything: Universal cross-modality image matching with large-scale pre-training. *arXiv:2501.07556 [cs.CV]*, 2025. DOI: 10.48550/arXiv.2501.07556

Contacts: Fateev Aleksey Sergeevich,
fateev.as@mipt.ru

Article received by the editorial office on 19.03.2026

INTRODUCTION

Cross-modal image matching, aimed at establishing spatial correspondences between images acquired by different sensor systems and under significantly different shooting conditions, is a key task in computer vision. In the context of remote sensing and autonomous systems, this task is particularly important, as accurate image matching serves as the basis for georeferencing, navigation, scene reconstruction, and real-time decision-making.

In particular, matching images obtained from unmanned aerial vehicles (UAVs) and satellites is an extremely complex cross-modal task. Images from UAV and satellite cameras differ fundamentally in scale, viewing angle, spatial resolution, and scene geometric structure. Satellite images are typically generated in an orthographic projection and cover large areas, whereas UAV images are characterized by a limited field of view, pronounced perspective distortions, and a high variability of perspectives. These differences significantly complicate the establishment of stable correspondences between images.

Modern UAV systems [1] are equipped with cameras with high frame rates, which leads to the formation of dense video sequences and significant streams of visual data. Meanwhile, methods for cross-modal matching of camera and satellite images are typically based on computationally complex models, including deep neural networks, dense [2] or semi-dense [3] matching algorithms, and geometric normalization procedures. As a result, the throughput of such models is significantly lower than the data rate from the sensors, making real-time processing of all frames virtually impossible. Under limited computing resources, the system is forced to partially process the input stream, which increases the likelihood of missing frames that are the most informative and provide optimal conditions for cross-modal matching.

Under such conditions, frame pre-selection becomes especially important. Its goal is to reduce the input data stream by eliminating frames that are unsuitable for subsequent cross-modal matching. Unlike classical video selection tasks, this requires selecting not just visually high-quality images, but frames that are highly informative, specifically for successful matching with satellite images.

The most common approaches to frame selection rely on general image quality metrics such as sharpness, contrast, or frequency-domain energy [4–6]. Such criteria effectively identify technically defective frames, such as blurry or noisy images. However, in cross-modal matching tasks, such metrics prove insufficient. Visually clear and contrasting UAV frames do not guarantee the presence of stable structural features consistent with the geometry of satellite images and,

therefore, do not necessarily lead to successful matching [6, 7].

The main reason for this is that classical quality metrics primarily evaluate low-level visual characteristics, without considering the structural and geometric compatibility of images of different modalities. In particular, fine texture details, shadows, moving objects, and localized illumination changes can significantly enhance the visual "quality" of a frame, but they do not contribute useful information for matching with satellite images dominated by large-scale and stable geometric structures such as road networks, buildings, or landscape boundaries.

Therefore, there is a need to develop frame pre-selection methods that focus not on the subjective or visual quality of an image but on assessing its potential suitability for cross-modal matching. Such methods should consider the structural information content (informativeness) of frames, their robustness to changes in scale and perspective, and the ability to maintain a consistent spatial organization of the scene between UAV and satellite images.

This paper proposes an approach to preselecting UAV camera frames aimed at improving the efficiency of cross-modal matching with satellite images under conditions of limited computational resources [8, 9]. Unlike existing methods focused on general image quality metrics, the proposed method focuses on analyzing the structural characteristics of the image relevant to the matching task. This significantly reduces the number of frames processed without losing georeferencing accuracy and improves the practical applicability of cross-modal matching systems in real-world scenarios.

PROBLEM STATEMENT

Let $F = I_{t(t=1)}^T$ be a video sequence obtained from an unmanned aerial vehicle's onboard camera at a high frame rate, where T denotes the total number of frames in the time interval under consideration, and I_s is a satellite image of the corresponding territory.

Due to limited computing resources, processing all T frames of a video sequence for cross-modal matching is practically impossible. Therefore, it is necessary to select a subset of frames

$$F' \subset F, |F'| \ll T,$$

that are most suitable for subsequent matching of UAV images with satellite data.

Unlike traditional frame selection methods based on visual image quality metrics, this paper addresses the task of assessing the structural informativeness of a frame, reflecting its ability to provide stable geometric correspondences in a cross-modal matching task.

Description of the transform

A two-dimensional discrete wavelet transform [10, 11] is used to analyze the structural characteristics of an image at different spatial scales. It provides joint spatial-frequency localization of features.

Let $I \in R^{H \times W}$ be the input image. The two-dimensional discrete wavelet transform is constructed by sequentially applying one-dimensional filters to the rows and columns of the image. Let g and h denote the high-pass and low-pass analysis filters, respectively. Then, the wavelet decomposition operations can be written as

$$L = (I * h) \downarrow 2, \quad G = (I * g) \downarrow 2,$$

where $(*)$ denotes convolution, and $\downarrow 2$ denotes down-sampling.

After sequential filtering by rows and columns, four sub-bands are formed:

$$(A_l, H_l, V_l, D_l),$$

where A_l is the approximating sub-band (low frequencies in both directions), and H_l , V_l and D_l represent horizontal, vertical, and diagonal image details, respectively.

The approximating sub-band A_l is recursively used to construct the next decomposition level, resulting in a hierarchical multi-scale representation of the image:

$$I \rightarrow \{A_L, H_L, V_L, D_L, \dots, H_1, V_1, D_1\},$$

where L denotes the number of decomposition levels.

This representation allows for the extraction of structural components of the image at different spatial scales, making wavelet decomposition a convenient tool for analyzing geometrically stable scene elements such as edges, texture patterns, and directional structures. An example of wavelet decomposition of a synthetic image is shown in Fig. 1. An artificially generated scene containing horizontal, vertical, and diagonal linear structures is used as input. The use of a synthetic image allows us to clearly demonstrate the directional selectivity of the 2D wavelet transform under controlled scene geometry, eliminating the influence of textural and photometric factors typical of real data.

Fig. 1. An example of 2D wavelet decomposition of a synthetic image.

a — original image; б — diagonal first-level detail coefficients D_1 ; в — horizontal first-level coefficients H_1 ; г — vertical first-level coefficients V_1

As can be seen from Figure 1, the first-level decomposition coefficients distinguish structures according to their orientation: horizontal components are concentrated in the subrange H_L , vertical components in V_L , and diagonal components in D_L . The approximating component preserves the overall geometry of the scene. This example illustrates the ability of the wavelet transform to perform spatially localized and directionally sensitive image analysis, which is a key property when assessing the structural informativeness of frames for comparison.

Unlike the Fourier transform [12], which has global frequency localization, the wavelet transform preserves the spatial localization of structural elements. This is especially important for images obtained from a UAV camera, which often contain distinct object boundaries, linear structures (roads, buildings, field boundaries), and textural elements of varying spatial scales. In a number of studies, simpler local metrics such as image entropy [13], gradient metrics [14], the Laplacian-based sharpness measure [15], and characteristics based on the structure tensor [16] are used to assess the informativeness of images. These methods reflect individual aspects of local contrast or statistical diversity of an image but do not take into account the interrelationships of structural elements at different spatial scales. The use of wavelet decomposition allows us to analyze the energy distribution of structural components simultaneously in the spatial and frequency domains, which makes this approach more suitable for identifying geometrically stable scene structures that are important for the tasks of establishing inter-image correspondences.

Assessing structural informativeness in the wavelet domain

Let an L -level 2D wavelet decomposition be performed on image I , yielding approximating and detailing subbands

$$(A_l, H_l, V_l, D_l), \quad l = 1, \dots, L.$$

The subbands H_l , V_l , and D_l are matrices of detail coefficients at level l , corresponding to horizontal, vertical, and diagonal image structures.

It is assumed that geometrically stable scene structures manifest themselves as local energy maxima in detailing subbands at several scales. For quantitative assessment, a normalized energy characteristic of level l is introduced:

$$E_l = \frac{1}{|H_l|} \sum_{i \in H_l} H_l(i)^2 + \frac{1}{|V_l|} \sum_{i \in V_l} V_l(i)^2 + \frac{1}{|D_l|} \sum_{i \in D_l} D_l(i)^2,$$

where $H_l(i)$ denotes the i -th coefficient of the corresponding subband, and $|\dots|$ is the number of coefficients in the subband.

Using the square of the coefficients corresponds to an L_2 norm signal energy estimate:

$$\|H_l\|_2^2 = \sum_i H_l(i)^2,$$

which is consistent with the energy interpretation of the wavelet transform as an orthonormal decomposition. Using orthonormal wavelets preserves energy, allowing us to construe E_l as the fraction of the image's structural energy localized at scale l .

Normalization by the number of coefficients $|H_l|$, $|V_l|$, $|D_l|$ ensures invariance of the estimate to the subrange size and allows for correct comparison of decomposition levels of different dimensions.

Thus, the quantity E_l reflects the total intensity of directed boundaries and extended structures at scale l , suppressing the contribution of random noise and small-scale textures due to quadratic averaging.

Interscale consistency of structures

To improve the robustness of the structural informativeness estimate, a measure of interscale consistency is introduced, reflecting the degree to which the image's energy characteristics are preserved when transitioning between adjacent decomposition scales. For levels l and $l+1$, interscale consistency is defined as

$$C_l = 1 - \frac{|E_l - E_{l+1}|}{E_l + E_{l+1} + \varepsilon}$$

where $\varepsilon > 0$ is a small regularizing parameter ensuring numerical stability if $E_l = E_{l+1} = 0$.

The quantity C_l takes values in the range $0 \leq C_l \leq 1$ and characterizes the degree of similarity of the energy distribution at adjacent scales. Values of C_l close to unity correspond to cases where the energy characteristics remain virtually unchanged between scales, indicating the presence of geometrically stable and extended scene structures.

Conversely, small values of C_l indicate significant differences between E_l and E_{l+1} , this is typical of small-scale textures, noise, or local artifacts which energy rapidly decays with increasing scale.

Estimation function and frame selection strategy

To ensure comparability of the contributions of different scales, the energy characteristics are normalized by the total energy of the detailing subranges:

$$\tilde{E}_l = \frac{E_l}{\sum_{k=1}^L E_k + \varepsilon},$$

where $\varepsilon > 0$ is a small regularizing parameter that ensures numerical stability at low energy values.

This normalization eliminates the dependence of the estimate on the overall image brightness and ensures that the condition is met.

$$\sum_{l=1}^L \tilde{E}_l \approx 1.$$

The final estimate of the structural informativeness of a frame I is defined as a weighted combination of the normalized energy and cross-scale characteristics:

$$S(I) = \sum_{l=1}^L \alpha_l \tilde{E}_l + \sum_{l=1}^{L-1} \beta_l C_l,$$

where the coefficients α_l and β_l define the contribution of the corresponding scales and are selected empirically. The first sum reflects the distribution of structural energy across the decomposition scales, and the second one reflects the degree of cross-scale stability of the identified structures.

Video sequence frames are ranked by value $S(I_t)$, after which a subset forms.

$$F = \{I_t : S(I_t) \text{ is included in top } K\},$$

where K is the specified number of selected frames $\ll T$.

A simplified calculation scheme for the proposed estimate is shown in Fig. 2. The method's input is an intersection image captured by a UAV camera, which is subjected to a multi-level wavelet decomposition. For each level, the energy characteristics of the detailing coefficients E_l and the inter-scale consistency indices C_l are calculated. The resulting values are aggregated into a final structural informativeness estimate $S(I)$, that is used to rank the video sequence frames.

Fig. 2. Calculation scheme for the estimate of the structural informativeness of a frame $S(I)$ based on a three-level wavelet decomposition of the image

To analyze the contribution of individual components, we consider the decomposition of the final criterion value $S(I)$. In this example, the weighting coefficients are set equal to unity ($\alpha_l = 1, \beta_l = 1$), allowing us to directly evaluate the relative contribution of the energy and inter-level components of the criterion. The decomposition results are presented in Tab. 1.

Tab. 1. The contribution of individual components to the final assessment of structural informativeness $S(I)$

From the data presented in Tab. 1, it follows that the main contribution to the criterion value $S(I)$ comes from the energy component of the third decomposition level $\tilde{E}_3$, accounting for approximately 66.6% of the total evaluation value. This indicates that a significant portion of the structural information in the scene is represented at larger spatial scales. This energy distribution is typical for images of urban infrastructure. The scene under consideration contains distinct geometric structures—road lines, intersections, and city block boundaries—which are primarily evident at coarser levels of the wavelet decomposition.

The total contribution of the consistency components between C_l decomposition levels is approximately 24% of the final criterion value, indicating the presence of spatial structures that are consistently evident at adjacent decomposition levels.

If necessary, an additional constraint on the minimum time interval between selected frames can be introduced to ensure a variety of observation angles. It is worth noting that calculating the estimate of $S(I)$ does not require information about the satellite image and can be performed at the early stages of processing.

Computational complexity

The proposed method has a computational complexity of $O(HW)$ per frame, where H and W denote the height and width of the image in pixels, respectively. Thus, the value of HW corresponds to the total number of image pixels. With the standard pyramidal implementation of 2D wavelet decomposition and a fixed filter length, the algorithm complexity remains linear in the number of pixels.

Linearity is ensured by the fact that the total computational effort across all decomposition levels decreases geometrically, since each subsequent level is applied to an approximating subrange that is half the size for each coordinate. As a result, the number of operations at level l is proportional to $HW / 4^l$, and the total complexity across all levels remains on the order of $O(HW)$.

Thus, when processing a video sequence of T frames, the total computational complexity of the structural analysis stage is $O(T \cdot HW)$.

Compared to the subsequent stage of cross-modal image matching, the computational cost of wavelet analysis is negligible. In particular, for modern dense

matching models [17, 18] based on the construction of a global correlation matrix or an attention mechanism, the computational complexity can reach $O(N^2)$ in terms of the number of tokens or spatial positions. Here, N denotes the number of elements in the spatial representation of the image (feature points after discretization of the feature space). Since N is usually proportional to the number of spatial positions in the feature map, the quadratic dependence leads to significantly higher computational costs compared to the linear complexity of preliminary wavelet analysis.

DESCRIPTION OF EXPERIMENTAL DATA

Video images

The main goal of the experiments is to analyze the impact of frame preselection on matching accuracy, registration stability, and the computational efficiency of the system as a whole. The experimental study utilized open datasets containing images obtained from unmanned aerial vehicles (UAVs) and satellite images of the corresponding terrain. The majority of the experiments were conducted using the UAV-VisLoc dataset [19], which is designed for visual localization of UAVs using satellite images. This dataset contains pairs of images: UAV images and corresponding satellite images of the same area. The images are characterized by significant variations in scale, viewing angle, and structural composition of the scene, making the matching task significantly more challenging. To analyze the performance of the methods in the infrared range, images from the VisDrone-DroneVehicle dataset [20] were additionally used. This dataset contains images obtained from unmanned aerial vehicles in the visible and infrared ranges of the spectrum. Since the infrared images in this dataset are not precisely georeferenced, corresponding satellite fragments were selected empirically based on the visual similarity of the scene structure. To simulate the complex UAV flight scenario, associated with abrupt maneuvers and deterioration of image quality, artificial image degradation methods were applied to some frames. Specifically, blurring, backlighting, and additive noise were used to simulate the effects of camera movement, illumination changes, and sensor degradation. These transformations allow the experimental conditions to approximate real-world flight scenarios, which involve abrupt changes in the camera's field of view, geometric distortions, and reduced image contrast. The resulting experimental sample contains frames with significantly different structural informativeness. Some images contain prominent geometric scene elements, such as object boundaries, road networks, and contrasting textures, while others are characterized by weak structural definition or significant

distortions. This allows us to objectively evaluate the ability of the proposed preselection method to identify frames with the greatest potential for successful cross-modal matching with satellite images. Examples of frames from the experimental sample are shown in Fig. 3.

Fig. 3. Examples of frames from the experimental sample.
a, б — frames in the visible range; в, г — frames in the infrared range

Experimental evaluation protocol

To ensure a fair comparison, all frame preselection methods were evaluated using an identical computational protocol. Experiments were conducted using two cross-modal matching models: LoFTR [17] and RoMa [18]. Since the open datasets used do not contain explicit video sequences, artificial video sequences were created from the UAV-VisLoc and VisDrone-DroneVehicle dataset images for the experiments.

The evaluation was performed on ten video sequences, each containing 60 frames. Sequences 1–7 corresponded to images obtained in the visible spectrum and created based on data from the UAV-VisLoc dataset. Sequences 8–10 contained infrared images and were created based on the VisDrone-DroneVehicle dataset.

For each matching model, all frame preselection strategies were compared with a constant network architecture, fixed parameters, and identical model weights. This allowed us to eliminate the influence of the matching parameters and focus on analyzing the effectiveness of the preselection methods.

Geometric verification was performed using the RANSAC algorithm [21] with fixed parameters, including the maximum number of iterations and the projection error threshold τ . As a result, differences in the number of found inlayers* are determined solely by the quality of frame preselection, all other things being equal for matching and geometric filtering.

Evaluation metrics

Metric types

Let a sequence contain N frames:

$$\{(S_i, M_i)\}_{i=1}^N,$$

* Inlayers are understood as geometrically consistent correspondences that satisfy the chosen transformation model.

where S_i is the value of the structural indicator for the i -th frame, and M_i is the number of inlayers obtained after the geometric verification step.

To quantitatively evaluate the effectiveness of the proposed approach, two groups of metrics are used:

- 1) correlation metrics, which characterize the statistical relationship between S_i and M_i ;
- 2) selection efficiency metrics, which reflect the practical gain in selecting the top K frames.

Correlation metrics

This group of metrics is designed to analyze the statistical relationship between the value of a structural indicator S_i and the number of inlayers M_i . Two correlation coefficients are used in this study.

Linear relationships are assessed using the Pearson correlation coefficient:

$$r = \frac{\text{cov}(S_i, M_i)}{\sigma_S \sigma_M},$$

where $\text{cov}(\cdot)$ is the covariance, σ_S and σ_M are the standard deviations of the variables and respectively.

Monotonic relationships are assessed using the Spearman rank correlation coefficient:

$$\rho = \text{corr}_{\text{rank}}(S_i, M_i).$$

For each coefficient, a corresponding p -value is calculated, characterizing the statistical significance of the identified relationship.

As part of the statistical hypothesis test, the null hypothesis H_0 declaring that there is no correlation between the variables S_i and M_i , is considered. The p -value is defined as the probability of obtaining the observed value of the correlation coefficient, given that H_0 is true.

If $p < \alpha$ the null hypothesis is rejected, the identified relationship is considered statistically significant. The study uses the significance level $\alpha = 0.05$.

Frame selection performance metrics

This group of metrics is intended to evaluate the practical usefulness of the proposed criterion for selecting a subset of frames.

Top- K frame selection. Let us denote the set of k frames with the highest values of the indicator S_i as T_k :

$$T_k = \text{Top-}K_k(S).$$

The average number of inlayers for the selected frames is defined as:

$$\mu_{\text{top}} = \frac{1}{k} \sum_{i \in T_k} M_i$$

A random selection of k frames from a sequence is considered the baseline. Let R_k be a random subset of cardinality k . The estimate of the average number of inlayers for random selection is calculated using the Monte Carlo method:

$$\mu_{\text{rand}} = E_{R_k} \left[\frac{1}{k} \sum_{i \in R_k} M_i \right],$$

hereinafter, $\text{Random} = \mu_{\text{rand}}$.

The absolute gain from using the proposed criterion is defined as:

$$\Delta @ k = \mu_{\text{top}} - \mu_{\text{rand}}.$$

The relative gain compared to random selection is defined as:

$$\text{Ratio} = \frac{\mu_{\text{top}}}{\mu_{\text{rand}}}.$$

EXPERIMENTAL RESULTS

Tabular presentation of results

This section presents an analysis of the evaluation results for the proposed frame preselection criterion. Since the absolute number of inlayers M_i depends significantly on the matching model used, experiments were conducted for two matching architectures: ELoFTR and RoMa. The experiments used pre-trained model weights from MatchAnything [22]. These models are ELoFTR and RoMa architectures additionally adapted for the task of cross-modal image matching (UAV $\leftrightarrow$ satellite). During training, network parameters were optimized to account for differences in spectral characteristics, contrast, and scale between aerial and satellite data. Thus, the comparison is not between the original general-purpose models but between their specialized versions adapted for cross-modal matching.

The experiments were conducted on ten video sequences, each containing $N = 60$ frames. From each sequence, $k = 15$ of the most informative frames were selected according to the proposed structural criterion. The summary quantitative results of the experiments are presented in Tabs 2 and 3.

Tab. 2. Results of the evaluation of the quality of frame selection for the model based on the ELoFTR architecture

Tab. 3. Results of the evaluation of the quality of frame selection for the model based on the RoMa architecture

Data analysis in table 2

Table 2 contains results for the ELoFTR architecture, while Table 3 corresponds to the RoMa architecture. The tables list the values of the correlation metrics (r , ρ) and the practical selection efficiency indicators (Top- k , Random , $\Delta @ k$, Ratio) for each sequence. Correlation metrics characterize the degree of consistency between the structural indicator and the actual matching success, while the Top- k , $\Delta @ k$, Ratio metrics reflect the practical efficiency of personnel selection.

The statistical significance of the correlation coefficients was assessed using the corresponding p -values. Coefficients for which the condition $p < 0.05$ was met were considered significant.

As can be seen from Table 2, most sequences exhibit a high positive correlation between the value of the structural indicator and the number of inlayers. For nine out of ten sequences, the Pearson coefficient exceeds 0.95, indicating a strong linear relationship between the structural informativeness of a frame and the matching success rate. The exception is sequence 3 ($r = 0.420$, $\rho = 0.548$), where the correlation is noticeably weaker.

The high values of the correlation metrics indicate that the proposed structural indicator adequately reflects the probability of successfully matching a frame to a satellite image. In other words, frames with more pronounced structural organization at different scales, as a rule, provide a greater number of stable geometric matches after the verification stage. This confirms the hypothesis that the structural informativeness of an image is an important indicator of the success of cross-modal matching.

The practical effectiveness of the proposed selection criterion is confirmed by the Ratio values, which exceed one for all sequences. This means that frame selection based on the structural criterion consistently outperforms random selection. The relative gain ranges from 1.28 to 2.87. The most pronounced effect is observed in sequence 3 ($\text{Ratio} 2.87$), despite the relatively low values of the correlation metrics. This result indicates that even with weakened global correlation, the criterion is capable of selecting a small subset of frames containing the most informative structural elements of the scene. A likely reason for this behavior is the heterogeneity of the video sequence structure, where only a small fraction of frames contains geometrically stable elements (roads, building boun-

daries, object outlines) that ensure a high number of matches.

Even in sequences with moderate rank correlation values (e.g., sequences 5–10), the practical gain remains significant. This indicates that the proposed criterion not only performs the function of ranking frames but also effectively concentrates computing resources on the most informative fragments of the video sequence, which is especially important in conditions of limited computing resources of onboard systems.

Data analysis in table 3

Table 3 shows parallel results for the RoMa model. The absolute values of the number of inlayers in this configuration are significantly higher compared to the ELoFTR architecture, reflecting the higher density of matches found. This is consistent with the characteristics of the RoMa architecture, which is designed to obtain denser and more spatially distributed matches.

For most sequences, a positive correlation is observed between the structural index and the number of inlayers; however, the degree of this relationship varies significantly more than in the case of ELoFTR. The highest rank correlation is observed in sequences 3 and 4 ($\rho \approx$ from 0.88 to 0.91), while the correlation is virtually not observed in some sequences (e.g., sequence 5). This may be due to the fact that RoMa is less sensitive to local structural features of images and is capable of finding matches even in conditions of weakly expressed geometric structure.

The practical preselection gain for RoMa is, on average, more modest. *Ratio* values in most sequences range from 1.01 to 1.15, indicating a relatively small but consistent advantage of structural selection over random selection. The most pronounced effect is observed in sequence 3 (*Ratio* 1.73), where the structural criterion significantly increases the number of found inlayers. In sequences 5 and 8, the gain is virtually nonexistent, which may be due to the structural homogeneity of the scene or the low variability of frames within a sequence.

Comparing the two configurations allows us to make several important observations. First, for both models, the proposed structural criterion demonstrates a positive correlation with matching success, confirming its model-independent nature and transferability across different matching architectures. Second, for the ELoFTR configuration, a more pronounced gain from frame preselection is observed. This indicates a higher sensitivity of this architecture to the quality of the input image set and highlights the importance of the prefiltering step when using less dense matching methods.

For RoMa, the effect of preselection is less pronounced, which may be due to the architecture's greater robustness to variations in the structural con-

tent of frames. Thanks to its denser matching mechanism, RoMa is capable of extracting matches even from less informative images, making the effect of preselection weaker. However, the positive gain remains in most sequences and confirms the practical feasibility of using the structural criterion as a stage of preliminary filtering of video stream frames.

Comparison with basic image quality metrics

To evaluate the advantages of the proposed preselection criterion, it was compared with a number of widely used image content metrics. The following basic methods were considered: entropy, gradient metric, Laplacian metric, and structure tensor metric. These metrics were chosen because they reflect different aspects of a frame's visual content: entropy characterizes the overall diversity of brightness values, gradient and Laplacian characterize the severity of local brightness variations and contour structures, and the structure tensor characterizes the intensity and directional consistency of local gradients.

Each of these metrics was used as a ranking criterion for video sequence frames, as it was earlier proposed to the wavelet criterion. In other words, for each sequence, the frames were sorted in descending order of the corresponding indicator value, then a subset of the k best frames was selected, on which the success of matching with the satellite image was then assessed.

In all experiments, the same selection parameters were used for the comparative methods: the sequence size was $N = 60$ frames, and the number of selected frames was $k = 15$. Since the *Random* value is common to all methods within a single sequence, it is not provided in the comparison tables. For each method, the Pearson r and Spearman ρ correlation coefficients were reported, as well as the practical selection efficiency indicators: *Top- k* , $\Delta@k$ and *Ratio*.

The comparison was conducted separately for two matching architectures: ELoFTR and RoMa. This separate analysis is necessary, since the effectiveness of the preliminary selection may depend on the specifics of the matching model used. The comparison results are presented in the Appendix in Tabs II1 and II2. Table II1 contains the results for the ELoFTR architecture, and Table II2 contains the results for the RoMa architecture. This representation allows us to evaluate not only the absolute quality of the proposed criterion but also its robustness to various cross-modal matching schemes.

As follows from Table II1, for the ELoFTR architecture, the proposed wavelet criterion demonstrates the highest consistency with the number of inlayers in most of the sequences considered. In particular, in sequences 01, 02, and 04–08, it provides maximum or near-maximum values of the Pearson r coefficient, in-

dicating its high ability to reflect changes in matching success from frame to frame. In other words, the wavelet representation better than the other metrics considered describes the quantitative relationship between the structural salience of an image and the number of correct geometric matches.

However, in terms of the practical Top- k , $\Delta@k$ and *Ratio* indicators, the advantage of the wavelet criterion is more heterogeneous. In a number of sequences, simple gradient metrics (gradient, Laplacian, structure tensor) demonstrate comparable or even higher practical gains. This is especially noticeable in sequences 01, 06, and 08, where local measures of contrast and contour expression prove sufficiently informative for frame selection. Nevertheless, the wavelet criterion demonstrates a significant advantage in sequences 02, 03, and 04, where it provides the highest Top- k , $\Delta@k$ and *Ratio* values. Sequence 03 is the most indicative: while the *Ratio* value for all baseline methods is approximately 1.67, for the wavelet criterion it increases to 2.87. This indicates that multiscale analysis allows us to identify truly informative frames in a structurally heterogeneous scene, while simple local metrics are insufficiently selective.

It should also be noted that for a number of sequences, the gradient, Laplacian, and structure tensor yield virtually identical Top- k , $\Delta@k$ and *Ratio* values. This behavior is predictable, since all three metrics are based on the analysis of local gradients and largely reflect the same type of structural informativeness, namely, the intensity of edges and contours. In contrast, the proposed wavelet criterion takes into account the distribution of structural energy across multiple scales, which ensures more reliable frame distinguishing in complex scenes.

The results for the RoMa architecture, presented in Table II2, reveal a different picture. On the one hand, the wavelet criterion again demonstrates high correlation metric values in some sequences, particularly in sequences 03, 04, and 07. On the other hand, its advantage in practical terms becomes less stable. In many sequences, gradient metrics provide comparable or greater gains, and the differences between the methods are significantly smaller than in the case of ELoFTR.

The most striking result for RoMa is observed in sequence 03. While entropy, gradient, Laplacian, and structure tensor lead to a decrease in selection quality relative to random selection (*Ratio* ≤ 0.82), the wavelet criterion, on the contrary, provides a marked improvement (*Ratio* 1.73, $\rho = 0.99$). This demonstrates that in complex sequences, the wavelet-based approach is capable of identifying robust multiscale structures, which are also significant for more powerful matching architectures. However, in most other sequences, the gain of the wavelet criterion for RoMa is either small or inferior to simpler baseline metrics.

Comparing the two tables allows us to draw several general conclusions. First, the proposed wavelet criterion has model-independent utility: for both architectures, it demonstrates a positive relationship with matching success and, in some sequences, provides the best practical result. Second, its advantage is most pronounced for the ELoFTR architecture, i.e., for a model more sensitive to the quality of the preselected frame set. In this case, analyzing image structure at different scales indeed provides an advantage over simple local metrics. Third, for RoMa, the impact of preselection is less pronounced, which is likely due to the greater inherent robustness of this architecture to variations in the structural informativeness of frames.

Therefore, it makes sense to consider the proposed criterion not simply as an alternative to traditional image quality metrics but as a specialized multi-scale indicator specifically designed for the task of preselecting frames for subsequent cross-modal matching. Its main advantage is evident in complex scenes, where local contrast characteristics do not allow for the reliable distinguishing of truly informative frames from less useful ones.

CONCLUSION

This paper examines the problem of pre-selecting informative frames for cross-modal UAV and satellite image matching systems under conditions of limited computing resources. The high frequency of video stream generation in UAV visual navigation systems results in a significant volume of input data, while the computational complexity of modern matching algorithms limits the number of frames that can be processed in real time. Under these conditions, it is necessary to prioritize input images and select a frame subset that is most suitable for subsequent matching.

A structural frame selection criterion is proposed, based on the analysis of image structure at different scales and the consistency of oriented components. Unlike traditional general-purpose image quality metrics, the proposed indicator focuses on identifying geometrically stable structures that play a key role in establishing correspondences between images.

An experimental evaluation was conducted on sequences generated from the open UAV-VisLoc and VisDrone-DroneVehicle datasets for two cross-modal matching configurations: ELoFTR and RoMa. The results obtained demonstrate a robust positive correlation between the structural indicator value and the number of inlayers to confirm the link between the multi-scale structural informativeness of a frame and the success of the matching procedure.

It has been shown that pre-selecting frames based on the proposed criterion increases the number of correct matches compared to random frame selection. For

the ELoFTR configuration, the gain reaches 1.3 to 1.8 times, demonstrating the high sensitivity of this architecture to the quality of the input image data. For the RoMa model, the selection effect is weaker; however, a positive increase in the number of matches is also observed in most sequences. Comparison with a number of basic image information metrics showed that the proposed wavelet criterion provides a more robust correlation with matching success and demonstrates advantages in scenes with complex structural organization. This confirms the feasibility of using multi-scale analysis to assess frame suitability for visual navigation tasks. Thus, the proposed approach can be considered a computationally efficient frame pre-selection mechanism in UAV visual navigation systems. Its application allows for increased robustness of cross-modal matching without increasing computational costs at the matching stage.

Further research into adaptive frame selection strategies that take into account scene dynamics and motion parameters, as well as the integration of the proposed structural metric directly into the training of cross-modal matching models, appears promising.

APPENDIX

Tab. II1. Comparison of the proposed wavelet criterion with popular image quality metrics for the ELoFTR architecture (correlation metrics and practical frame selection efficiency indicators are provided for each sequence)

Tab. II2. Comparison of the proposed wavelet criterion with popular image quality metrics for the RoMa architecture (correlation metrics and practical frame selection efficiency indicators are provided for each sequence)