ОБРАБОТКА ИЗОБРАЖЕНИЙ, РАСПОЗНАВАНИЕ ОБРАЗОВ

О КОЛИЧЕСТВЕННОЙ ОЦЕНКЕ ЭФФЕКТИВНОСТИ АЛГОРИТМОВ АНАЛИЗА ИЗОБРАЖЕНИЙ

П.П. Кольцов, А.С. Осипов, А.С. Куцаев, А.А. Кравченко, Н.В. Котович, А.В. Захаров
Научно-исследовательский институт системных исследований РАН, Москва, Россия

Аннотация

Статья содержит краткий обзор основных подходов к сравнительной оценке эффективности алгоритмов анализа изображений. Рассмотрены эмпирические методики сравнительного исследования эффективности детекторов границ и алгоритмов сегментации изображений, проводится анализ используемых при этом количественных критериев их оценки. Описан ряд проблем, возникающих при использовании этих критериев. Изложена эмпирическая методика EDEM (на примере сравнительного тестирования детекторов границ), реализуемая в рамках разрабатываемой авторами программной среды PICASSO.

Ключевые слова: сравнительное исследование, анализ изображений, детекторы границ, сегментация изображений, мера эффективности, ground truth образ, нечёткие множества.

Цитирование: **Кольцов, П.П.** О количественной оценке эффективности алгоритмов анализа изображений / П.П. Кольцов, А.С. Осипов, А.С. Куцаев, А.А. Кравченко, Н.В. Котович, А.В. Захаров // Компьютерная оптика. – 2015. – Т. 39, № 4. – С. 542-556. – DOI: 10.18287/0134-2452-2015-39-4-542-556.

Введение

Известно, что сканирование и компьютеризированная обработка изображений начали проводиться в 1956 году в Национальном бюро стандартов США. В тот же период начали разрабатываться алгоритмы улучшения изображений [1]. Более чем полвека спустя были созданы тысячи разнообразных алгоритмов обработки изображений. Некоторые из них разрабатывались для отдельных специальных приложений (таких как, например, улучшение скрытых отпечатков пальцев), в то время как другие предполагали более разнообразные области применения, зачастую без явных предпочтений. Сфера действия этих алгоритмов весьма обширна: от автоматического выделения и разграничения областей интереса в задаче сегментации до повышения воспринимаемого качества картинки методами улучшения изображений. С началом развития компьютерной обработки изображений, как и в других компьютерных областях, тестирование разрабатываемых алгоритмов стало составной частью данного процесса. Цель тестирования двояка. Во-первых, оно предоставляет количественный или качественный метод оценки алгоритма. Во-вторых, оно даёт сравнительную оценку алгоритма относительно других алгоритмов по определённым критериям. Выбор критерия, используемого при анализе результатов, является непростой задачей. Чему, к примеру, отдавать предпочтение: аккуратности, устойчивости или чувствительности? Оценка эффективности в широком смысле есть мера требуемого поведения алгоритма, определяющая, достигнуты ли приемлемые уровни аккуратности, устойчивости и адаптивности его работы. Она позволяет выделить существенные свойства алгоритма, а также оценить его достоинства и ограничения. С данной оценкой тесно связан процесс анализа его ошибок и сбоев. Такой анализ прежде всего требует определения характеристик успеха. Особенно важен этот процесс на этапе разработки алгоритма, так как он позволяет вносить в него исправления и добавления. Здесь процесс анализа ошибок можно рассматривать как начальное тестирование.

Такое всестороннее тестирование в настоящее время не является общепринятой практикой. Отчасти это объясняется недостаточностью формального процесса, используемого при оценке эффективности: от выбора режимов тестирования до разработки мер эффективности. Кроме того, различным подходам к сравнительному исследованию алгоритмов обработки изображений в литературе за последние полвека уделялось относительно мало места. При этом следует учесть, что выбор методологии тестирования алгоритма зависит от специфики задачи, в которой он используется.

Как отмечалось в [2], целью оценки алгоритма является понимание его поведения на различных категориях изображений и помощь в выборе наилучших параметров алгоритма в различных ситуациях. На завершающей стадии оценивания требуется его сравнение со сходными алгоритмами с целью выработки практических рекомендаций по применению алгоритмов рассматриваемого семейства в той или иной ситуации. Оценка любого алгоритма обработки изображений зависит от нескольких факторов [3]:

- самого алгоритма,
- изображений, используемых для измерения эффективности алгоритма,
- параметров алгоритма, используемых в его тестировании,
- метода оценки эффективности.

Сложность в оценке алгоритма напрямую зависит от числа его параметров. Для оптимизации работы алгоритма требуется проводить их выбор, что само по себе непросто. Кроме того, на различных тестовых изображениях эти параметры могут меняться. Что касается влияния тестовых образов на оценку качества алгоритмов, то использование в тестировании относи-

тельно простых изображений часто даёт существенно лучший результат, чем использование изображений, моделирующих ситуации, сложные для распознавания.

В настоящее время не существует строгого регламента, характеризующего процесс оценки, однако здесь следует принимать во внимание следующий ряд факторов: протокол тестирования, режим тестирования, индикаторы эффективности, меры эффективности и базы тестовых образов [4]. Первый из них, протокол тестирования, относится к последовательному подходу, используемом в этом процессе. Следующий из вышеуказанных факторов, режим тестирования, определяет стратегию данного процесса. Здесь выделяют четыре основные категории тестирования. Первая из них, исчерпывающее тестирование, представляет собой грубый подход к тестированию, основанный на переборе всех тестовых изображений из базы данных. Этот подход зачастую является чрезмерным и может быть ограничен этапом верификации разрабатываемого алгоритма. Следующая категория, тестирование граничных значений (boundary value testing), оценивает работу алгоритма на определённой пользователем репрезентативной выборке изображений базы данных. Третья категория, случайное тестирование, основано на произвольном выборе тестовых образов. При таком тестировании могут возникать более разнообразные ситуации, чем при тестировании граничных значений, поскольку в последнем случае выбор изображений носит субъективный характер и может не учитывать всего многообразия случаев, возникающих на практике. К примеру, при проверке алгоритма распознавания злокачественных опухолей в первую очередь будут выбираться изображения, содержащие эти опухоли, хотя на практике чаще встречаются случаи, когда опухоли доброкачественные или вообще отсутствуют. Последняя категория, тестирование в наихудших случаях (worst case testing), включает в себя рассмотрение ситуаций, когда тестовый образ содержит редкие или необычные свойства.

Что касается индикаторов эффективности, то это характеристики (определение которых достаточно произвольно), определяющие качества алгоритма. Измерение их весьма трудно. К типичным индикаторам относятся:

- аккуратность: насколько хорошо алгоритм работает с тестовыми изображениями,
- устойчивость: способность алгоритма работать в различных условиях,
- чувствительность: как алгоритм реагирует на малые изменения входных данных,
- адаптивность: автоматическое подстраивание алгоритма под особенности разнообразных образов,
- надёжность: проверка близости результатов при повторении теста с теми же входными данными,
- эффективность: практическая применимость алгоритма (удобство, затратность, модернизируемость и т. п.).

Наконец, понятие базы тестовых образов определяет, какие изображения следует брать для тестирования. Здесь учитываются такие факторы, как разнообразие изображений, их уровень сложности и значимость при тестировании тех или иных классов алгоритмов (например, алгоритмов сегментации или выделения границ).

1. О методиках оценки эффективности алгоритмов

Как отмечалось выше, к настоящему времени созданы тысячи разнообразных алгоритмов обработки изображений. Многие из них имеют различные программные реализации, в том числе находящиеся в открытом доступе (например, это относится к известному алгоритму Canny выделения границ). В результате перед разработчиками систем компьютерного зрения встаёт непростая проблема выбора наиболее адекватных их задачам алгоритмов.

Тестирование алгоритмов для выявления наиболее адекватных решаемым задачам обработки изображений, в силу указанных выше причин, представляет собой процесс, не имеющий единой методики. Основные отличия методик, применяемых в сравнительных исследованиях алгоритмов, решающих одну и ту же задачу (выделение границ, сегментация, поиск текстурных областей и т. п.), друг от друга:

- разные наборы тестовых изображений, отличающиеся как по типу изображений (реальные или синтезированные), так и по размеру, количеству, источникам (оригинальные изображения или из общедоступных БД) и т.д.;
- разные процедуры выбора оптимальных параметров алгоритмов;
- разные критерии оценки качества алгоритмов (количественные или качественные, использующие эталонное изображение или нет).

К настоящему времени было сделано несколько попыток классифицировать эти методики. Так, в [5] была предложена классификация методик сравнительного исследования алгоритмов сегментации изображений, в соответствии с которой методики оценки делятся на:

1) субъективные,
2) объективные,
 2.1) системные,
 2.2) прямые,
  2.2.1) аналитические,
  2.2.2) эмпирические,
   2.2.2.1) контролируемые,
   2.2.2.2) неконтролируемые (автоматические).

В принципе, данная классификация пригодна и для методик оценки других видов алгоритмов анализа изображений (например, детекторов границ).

Субъективные (они же визуальные) – наиболее широко используемые методики оценки. Их основной недостаток, собственно, отражён в названии этого класса – оценка качества даётся человеком, поэтому у разных экспертов эта оценка может кардинально отличаться.

Объективные методики, не использующие визуальные оценки, подразделяются на системные и прямые.

Системные методики дают оценку алгоритму на основе конечных результатов работы всей системы распознавания изображений. В качестве примера можно привести оценку работы разных детекторов границ на основе результатов распознавания объектов, выделенных на изображении [6]. Такая оценка не обязательно говорит о качестве работы алгоритма, а может просто указывать на более подходящий результат для дальнейшей обработки.

Прямые методики, имеющие дело непосредственно с самим исследуемым алгоритмом или с результатами его работы, подразделяются на аналитические и эмпирические.

Аналитические методики рассматривают алгоритм независимо от его выхода [7]. Изучаются такие свойства алгоритма, как стратегия реализации главной цели, сложность, возможность распараллеливания, ресурсоёмкость и т.п. Эти свойства не имеют прямого отношения к качеству работы алгоритма. Аналитические методики, рассматриваемые в литературе, имеют дело в основном с задачами специального вида (см., например, [8]).

Эмпирические методики, напротив, оценивают не сам алгоритм, а результаты его работы на некотором наборе тестовых изображений. Они подразделяются на контролируемые и неконтролируемые (автоматические).

Контролируемые (supervised) методики часто называются в англоязычной литературе discrepancy methods (см., например, [9]), что, возможно, более точно соответствует их сути, так как они используют для оценки количественные меры различия результата работы алгоритма с некоторым эталонным результатом – ground truth изображением. Последние, зачастую созданные искусственно, содержат идеальные с точки зрения экспериментатора результаты. Например, если исследуются детекторы границ, каждому тестовому изображению соответствует ground truth образ, содержащий идеальные границы. Возможна ситуация, когда для каждого алгоритма исследуется несколько его свойств, тогда каждому тестовому изображению может соответствовать сразу несколько ground truth образов (см. также [23]). Такие методики дают очень хорошую оценку. Но создание таких эталонных образов для многих тестовых изображений зачастую требует больших трудозатрат (например, при создании эталонной сегментации изображений реального мира для исследования алгоритмов сегментации) и вносит элемент субъективности.

В автоматических (unsupervised) методиках (другое название – goodness methods [9]) производится количественная оценка некоторых желаемых свойств изображения, обработанного с помощью исследуемого алгоритма, на основе чего делается вывод о качестве последнего. Они не требуют наличия эталонных изображений, что, возможно, является их основным достоинством. Это свойство позволяет осуществлять контроль и самообучение в системах реального времени.

Одним из ключевых элементов методики сравнительного тестирования является используемый критерий оценки качества работы алгоритма (соответствующие английские варианты этого термина: evaluation criterion, performance criterion, performance metric, performance measure, performance index [10]).

Данная статья посвящена эмпирическим методикам оценки. Основное внимание при этом уделяется контролируемым методикам. В частности, в следующем разделе рассматриваются основные количественные меры оценки эффективности детекторов границ, анализируются особенности их применения и недостатки. Далее анализируются основные критерии оценки качества работы алгоритмов сегментации и соответствующие им классы мер. В следующем разделе рассматривается разрабатываемая нами методология EDEM сравнительного исследования алгоритмов анализа изображений (на примере детекторов границ), реализованная в программной системе PICASSO.

Перед тем как рассматривать предложенные в литературе количественные меры для оценки качества работы того или иного класса алгоритмов анализа изображений, весьма желательно определить, каким же требованиям должно соответствовать изображение, обработанное с использованием исследуемого алгоритма. Например, применительно к исследованию алгоритмов сегментации изображений такие требования были сформулированы в [11] (см. ниже). В большинстве случаев именно этим качественным признакам исследователи и пытаются сопоставить некоторые количественные критерии оценки качества тестируемых алгоритмов.

2. Количественные оценки эффективности детекторов границ

Принято считать, что основные требования к детектору границ были впервые сформулированы J. Canny в классической работе [12]. Автору удалось сначала сформулировать содержательные требования к детектору, выразить их в виде некоторой оптимизационной задачи и, наконец, решить эту задачу. Canny потребовал, чтобы детектор границ удовлетворял следующим трём критериям эффективности:

1. Хорошее отношение сигнала к шуму. Содержательно это означает, что детектор должен выделять все истинные границы и при этом не выделять ложных.

2. Хорошая локализация – точки, отмеченные как границы, должны располагаться как можно ближе к истинному положению границ.

3. Единственность отклика на границу (с одной стороны, это требование содержится в первом критерии, однако математическая формулировка первого критерия не обеспечивает выполнения данного пункта).

J. Canny изучил математическую проблему получения фильтра границ, оптимального по данным критериям. Он показал, что искомый фильтр является суммой четырёх экспонент, и может быть хорошо

приближен первой производной Гауссиана. Хотя данная работа была выполнена на заре компьютерного зрения, детектор границ Canny до сих пор является одним из лучших.

Требования, сформулированные Canny в отношении «хорошего» детектора границ, можно считать теми признаками, упомянутыми в предыдущем разделе, с которыми сопоставляются количественные оценки эффективности детекторов границ. В терминах эмпирической контролируемой методики сравнительного тестирования эти требования означают, что основными характеристиками хорошего детектора границ является высокий процент правильно выделенных граничных пикселей (высокий уровень детектирования) и высокая степень локализации (близость выделенных пикселей к соответствующим им на ground truth изображении). При этом, как было отмечено Canny, имеет место своего рода принцип неопределённости между высоким уровнем детектирования и высокой локализацией. Отчасти этим объясняется то, что к настоящему времени не удалось создать меру, одинаково эффективно оценивающую эти две характеристики. Соответственно, в ряде работ, посвящённых исследованию мер эффективности для детекторов границ (см. [13] и цитируемую там литературу), рассматриваемые меры подразделялись на меры оценки качества детектирования, или статистические (detection performance, or “statistical” measures), и меры оценки локализации (localization performance, or “distance” measures).

Приведём пример нескольких часто встречающихся мер оценки качества детектирования (подробнее см. [13]). Итак, пусть имеются: X – растр изображения, состоящий из N пикселей, B – результат работы оцениваемого детектора границ (образ, состоящий из граничных точек) и A – соответствующий ground truth образ. Тогда ошибка 1-го рода определяется как:

$$\alpha(A,B)=\frac{n(B\setminus A)}{n(X\setminus A)},$$

где $n(.)$ – число пикселей в соответствующем множестве, т. е. как отношение неправильно выделенных граничных пикселей к общему числу пикселей, не являющихся граничными.

Ошибка второго рода определяется как

$$\beta(A,B)=\frac{n(A\setminus B)}{n(A)},$$

т. е. как отношение не выделенных граничных пикселей к общему числу граничных пикселей.

Также на практике весьма часто встречаются такие меры как чувствительность (sensitivity):

$$Se=\frac{n(B\cap A)}{n(A)}=1-\beta,$$

т. е. это отношение правильно выделенных граничных пикселей к общему числу граничных пикселей, а также специфичность (specificity):

$$Sp=\frac{n(X/B\cap A)}{n(X/A)}=1-\alpha.$$

Указанные меры изначально нашли своё применение в медицинской статистике в анализе диагностики различных заболеваний.

К мерам оценки качества детектирования можно отнести и среднеквадратическую Евклидову метрику, применяемую при сравнении двух полутоновых изображений. Сюда же относятся эквивалентные (в данной ситуации) отношения сигнала к шуму: пиковое и среднеквадратическое.

Указанные меры оценки качества детектирования имеют большое практическое применение, при этом неоднократно отмечались их недостатки. Наиболее слабым их местом является то, что различия между изображениями A и B определяются по общему числу расхождений между ними, безотносительно к образу, который эти изображения представляет. Так, искажения, затрагивающие относительно незначительное число пикселей, но существенно меняющие форму изображаемого объекта (небольшие удаления линий, заполнения маленьких дырок и т. п.) дадут хорошие значения данных мер. Рассмотрим в качестве примера тестовое изображение на рис. 1*а* и соответствующий ground truth образ на рис. 1*б*. Пусть на изображении тестируются два детектора границ, результаты их работы приведены на рис. 1*в* и рис. 1*г* соответственно.

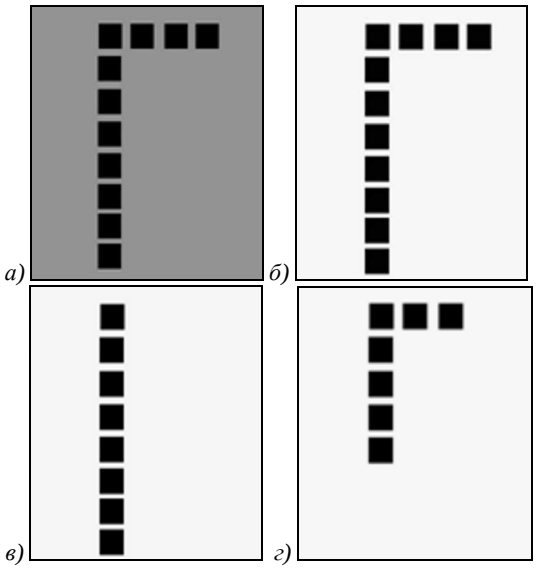


Рис. 1. а) Тестовое изображение, б) соответствующий ground truth образ, в) и г) результаты обработки изображения а) двумя детекторами границ

Здесь для обоих детекторов значения ошибки первого рода и специфичности одинаковы и равны соответственно 0 и 1 (все граничные пиксели выделены правильно). Значения же ошибки второго рода и чувствительности для первого детектора равны 0,27 и 0,73, а для второго равны соответственно 0,36 и 0,64. Таким образом, если руководствоваться только данными этих мер, на рассматриваемом изображении первый детектор даёт результат лучше, чем второй, что противоречит визуальной оценке (и здравому смыслу).

К проблеме проверки сохранения формы границ посредством применения статистических мер примыкает проблема правильного выбора порога для определения соответствия пикселей на двух изображениях. Так, небольшой сдвиг в оцениваемой карте границ относительно ground truth изображения, затрагивающий большое число пикселей, но не меняющий формы образа (т. е., к примеру, на оцениваемом изображении то же яблоко, что и на ground truth, слегка сдвинутое по отношению к последнему), может привести к плохим значениям оценки качества детектирования. При практическом оценивании детекторов границ с помощью этих мер следует принимать во внимание указанные недостатки. В частности, это относится к выделению границ, использующих размытие зашумлённых изображений на этапе предобработки.

Что касается мер оценки локализации, к ним можно отнести среднеквадратическую ошибку расстояния (mean square error distance):

$$e(A,B)=\frac{1}{n(B)}\sum_{x\in B} d(x,A)^2,$$

где $d(X,A) = \inf \rho(x,a)$, $a \in A$, а $\rho(.,.)$ в рассматриваемой дискретной ситуации представляет собой метрику кратчайшего пути (shortest path length metric, см. [14] и приведённые там ссылки), а также часто встречающуюся в литературе метрику Пратта (Pratt's figure of merit):

$$FOM(A,B)=\frac{1}{\max\{n(A),n(B)\}}\sum_{x\in B}\frac{1}{1+kd(x,A)^2},$$

где k – коэффициент масштабирования, обычно полагаемый равным 1/9, а метрика кратчайшего пути нормируется таким образом, что наименьшее ненулевое расстояние между соседними пикселями равно единице. Очевидно, $0 < FOM(A,B) \le 1$ и равно единице тогда и только тогда, когда $A=B$.

К данному разряду мер оценки локализации можно отнести и метрику Хаусдорфа:

$$H(A,B)=\max\{\sup_{a\in A} d(a,B), \sup_{b\in B} d(b,A)\}.$$

Хотя классическая версия данной метрики обладает желательными для использования в оценке цифровых изображений топологическими свойствами, на практике она используется сравнительно редко в силу своей высокой чувствительности к шуму и даже к изменениям в один пиксель на оцениваемом изображении. Наиболее часто используемой из перечисленных мер является метрика Пратта.

Как и меры оценки качества детектирования, меры оценки локализации имеют ряд недостатков. Они могут быть нечувствительными к ошибкам второго рода. Например, если $B \in A$, то значения $e = 0$, а $FOM(A,B) = n(B)/n(A) = 1-\beta(A,B)$, т.е. значение $FOM(A,B)$ совпадает со значением специфичности и не несёт никакой новой информации. Среднеквадратическая метрика и особенно метрика Хаусдорфа сильно чувствительны к фоновому шуму. Что касается метрики Пратта, в ряде случаев возникали ситуации, когда она давала высокие значения при наличии разрывов границ (дырок) или когда выделенная граница осциллирует вокруг своего истинного положения (на ground truth изображении) (см. [13]).

Также в работе [14] был приведён получивший известность противоречащий здравому смыслу пример поведения метрики Пратта (т.н. пример Peli-Malah) (рис. 2*а-в)*.

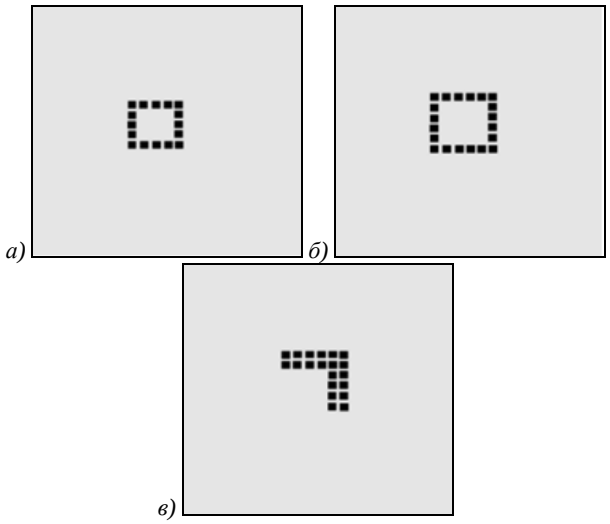


Рис.2. Пример Peli-Malah: а) ground truth образ (квадрат 5×5), б) и в) оцениваемые образы. Значение FOM в случаях б) и в) одинаково и равно 0,941

Авторами примера были предъявлены два изображения, приведённые на рис. 2*б, в*, на которых значения *FOM* будут одинаковыми, если в качестве ground truth образа взять квадрат размером 5×5 пикселей, изображенный на рис 2*а*. Подобные примеры означают, что метрика Пратта позволяет в рамках выбранного порога соответствовать нескольким выделенным граничным пикселям одному и тому же пикселю на ground truth изображении (в случае отсутствия взаимно однозначного соответствия пикселей её значение может быть таким же, как и в случае наличия такого соответствия). Заметим, что и значения приведённых выше мер оценки качества детектирования будут также одинаковы. В работе [14] была предложена мера оценки локализации Δ_w^p, представляющая собой L^p модификацию метрики Хаусдорфа:

$$\Delta_w^p=\frac{1}{n(X)}\left(\sum_{x\in X}|\, w(d(x,A))-w(d(x,B))\,|^p\right)^{1/p},$$

$$1\le p\le\infty,$$

где в качестве w, как правило, берётся т.н. функция отсечения (cutoff transform): $w(t)=\min\{t,c\}$ для некоторого фиксированного $c>0$. В отличие от метрики Хаусдорфа, она более устойчива к возмущениям в один или несколько пикселей (максимум в разности расстояний заменяется на L^p среднее). Ряд экспериментов показал её более адекватное поведение в сравнении с метрикой Пратта. Так в примере Peli-

Malah на рис. 2*б*,*в* значения Δ^2 при $c = 5$ были 0,323 и 0,512 соответственно. В то же время метрика Пратта продемонстрировала более устойчивое поведение к малым возмущениям границы.

В последние годы для оценки эффективности детекторов границ был предложен ряд новых статистических мер оценки локализации. Учитывается взаимно однозначное соответствие пикселей на изображении, используется информация о силе края (edge strength) и т. п. [13]. Данные меры также не свободны от недостатков. Заметим, что в отличие от вышеприведённых мер, некоторые из них имеют сложную программную реализацию и их вычисление представляет собой трудоёмкий процесс, например, проводится поиск соответствующих пар пикселей по минимальному значению функции затрат, для чего используются алгоритмы из теории графов. Тем самым задача создания метрик эффективности детекторов границ, адекватных по своим результатам и несложных в реализации, сохраняет свою актуальность.

Что касается изображений, используемых в эмпирических контролируемых методиках оценки детекторов границ, следует отметить, что в ряде работ использовались создаваемые искусственно несложные ground truth или брались несложные реальные изображения, содержащие легко идентифицируемые границы. На практике нередко встречаются достаточно трудные для распознавания ситуации, что ограничивает область применимости методик, предложенных в данных работах. Остаётся недостаточно исследованным вопрос и о полноте набора тестовых изображений, используемых в методологии с использованием ground truth образов.

Также, в ряде работ (см., например, [13] и содержащиеся там ссылки) используются трёхзначные ground truth образы, где каждый пиксель изображения относится к одному из трёх классов: граница, фон, не важно (попадание в этот класс на оценки не влияет). Это упрощает искусственное создание ground truth образов, соответствующих реальным изображениям (можно, например, отнести к третьему классу все текстурные области на изображении), а также делает количественные оценки более содержательными. Заметим, что в полутоновых изображениях имеет место неопределённость, существующая в локализации положения границы, отделяющей объект от фона. Особенно это характерно для размытых изображений. Это является одной из трудностей в создании двузначных ground truth образов, содержащих эталонные карты границ, и делает перспективным использование элементов теории нечётких множеств в исследовании эффективности детекторов границ (ниже мы подробнее остановимся на этом вопросе).

3. О критериях оценки качества сегментации

Прежде чем рассматривать известные к настоящему времени количественные критерии оценки сегментации, необходимо определить требования, которым должно удовлетворять сегментированное изображение. Как отмечено выше, признаки «хорошей» сегментации были сформулированы в [11]:

– сегменты должны быть однородными по некоторым характеристикам (например, по яркости или текстуре);

– соседние сегменты должны значительно отличаться по этим характеристикам;

– внутри сегмента не должно быть большого количества мелких «дырок»;

– границы сегментов должны быть гладкими и иметь точную пространственную локализацию.

В большинстве случаев именно этим качественным признакам исследователи и пытаются сопоставить некоторые количественные критерии оценки сегментации изображения.

В литературе выделяются два основных подхода к сегментации изображений:

1) разделение изображения контурами на области со сходными характеристиками (в английской терминологии – edge-based methods [10] (также используются термины boundary-based и contour-based methods));
2) объединение пикселей изображения в группы на основе близости некоторых количественных признаков (region-based methods [10]).

Для оценки результатов работы алгоритмов, относящихся к первой группе, используются в основном те же критерии, что и для детекторов границ (см. предыдущий раздел). Ниже мы рассмотрим количественные критерии оценки, используемые для алгоритмов сегментации, относящихся ко второй группе.

Самая простая и естественная мера качества сегментации, которую сразу же начали использовать исследователи, занимавшиеся сегментацией изображений, – это процент неправильно классифицированных пикселей. Очевидно, она относится к разряду статистических мер оценки качества сегментации. Однако у этого критерия имеются явные недостатки:

– иногда результаты сегментации, явно лучшие с точки зрения экспертов, имели более высокий процент ошибочно классифицированных пикселей (недостаток, характерный для статистических мер, см. предыдущий раздел);

– не учитывалось расположение ошибочных пикселей относительно соответствующего сегмента – очевидно, что ошибка на границе и ошибка в центре сегмента должны штрафоваться по-разному;

– не учитывалось различие в важности отдельных участков изображения для сегментации – ошибки для разных его сегментов должны иметь разный вес;

– отсутствовала информация о том, какой класс пикселей вносил наибольшую ошибку.

Для решения последних двух проблем в [15] было предложено два критерия, являющихся обобщением для случая нескольких классов ошибок первого и второго рода (см. выше). Оба критерия основаны на построении матрицы неточностей (confusion matrix). Столбец этой матрицы соответствует классу, к которому пиксели принадлежат на самом деле, а строка –

классу, к которому пиксели отнесены при сегментации (т. е. правильно классифицированные пиксели относятся к элементам матрицы, находящимся на главной диагонали, неправильно классифицированные – ко всем остальным).

Первый из предложенных критериев – процентное отношение неправильно классифицированных пикселей данного k-го класса к общему количеству пикселей этого класса на эталонном изображении:

$$M_1^k = \left(\left[\left(\sum_{i=1}^{n} C_{ik} \right) - C_{kk} \right] \Big/ \sum_{i=1}^{n} C_{ik} \right) \times 100,$$

где n – количество классов, C_{kk} – количество правильно классифицированных пикселей k-го класса, $\sum_{i=1}^{n} C_{ik}$ – количество пикселей, в действительности принадлежащих к k-му классу (k-й столбец матрицы неточностей).

Второй критерий – это процентное отношение пикселей, ошибочно причисленных к данному k-му классу, к общему количеству пикселей других классов на эталонном изображении:

$$M_2^k = \frac{\left(\sum_{i=1}^{n} C_{ki} \right) - C_{kk}}{\left(\sum_{i=1}^{n} \sum_{k=1}^{n} C_{ik} \right) - \sum_{i=1}^{n} C_{ik}} \times 100,$$

где $\sum_{i=1}^{n} C_{ki}$ – количество пикселей, отнесённых к k-му классу при сегментации (k-я строка матрицы ошибок), $\sum_{i=1}^{n} \sum_{k=1}^{n} C_{ik}$ – число пикселей на изображении.

Таким образом, при наличии n сегментов изображения получаем $2n$ критериев M_1^k, M_2^k, k=1,2…n, позволяющих проанализировать вклад каждого сегмента в общую ошибку. Кроме того, элементы матрицы неточностей могут быть взвешены с целью учесть разную значимость ошибок для разных сегментов изображения. Однако явной процедуры построения этих весов предложено не было. Заметим, что учёт важности отдельных участков изображений для сегментации может быть формализован с помощью элементов нечёткой логики посредством построения нечётких ground truth образов и использования нечётких мер эффективности (см. следующий пункт).

Другая статистическая мера оценки качества сегментации, основанная на подсчёте неправильно классифицированных пикселей и использующая Байесовский подход, была предложена в [16]. В этой работе вычисляются вероятности того, что случайно выбранный пиксель на отсегментированном изображении принадлежит объекту и, соответственно, фону. Используя эти вероятностные формулы, выводится вероятность ошибки сегментации всего изображения:

$$p(err) = p(o)\,p(b \mid o) + p(b)\,p(o \mid b),$$

где $p(o)$, $p(b)$ – априорные вероятности того, что случайным образом выбранный пиксель исходного изображения принадлежит объекту или, соответственно, фону. Они вычисляются по ground truth образу как отношение суммы пикселей (площади) объекта и, соответственно, площади фона к площади всего изображения; $p(o \mid b)$ – вероятность того, что пиксель, принадлежащий фону, будет при сегментации ошибочно отнесён к объекту. Она вычисляется как отношение суммы пикселей фона, ошибочно отнесённых к объекту на сегментированном изображении к сумме пикселей (площади) фона на эталонном изображении; $p(b \mid o)$ – вероятность того, что пиксель, принадлежащий объекту, будет ошибочно отнесён к фону, вычисляемая аналогичным образом. В дальнейшем эта формула была обобщена на случай представления объекта в виде произвольного числа сегментов.

Приведённые выше меры, основанные на подсчёте неправильно классифицированных пикселей, не учитывают взаимного расположения такого пикселя и сегмента, к которому он был ошибочно отнесён. Понятно, что чем больше расстояние между ними на эталонном изображении, тем значительнее такая ошибка должна штрафоваться.

В упоминавшейся выше работе [15] была предложена мера, основанная на таком подходе:

$$\varepsilon = \left(\sqrt{\sum_{i=1}^{N} d_i^2} \Big/ A \right) \times 100,$$

где N – количество ошибочно классифицированных пикселей, A – общее количество пикселей в изображении, d_i – евклидово расстояние между i-м ошибочно классифицированным пикселем x и ближайшим пикселем y, действительно относящимся к данному классу.

Очевидно, данная мера представляет с собой аналог мер оценки локализации, используемых для тестирования детекторов границ. Таким образом, в работе [15] для тестирования алгоритмов сегментации используются меры двух классов: статистические и меры оценки локализации. Этот подход соответствует и нашей методологии тестирования, представленной в следующем разделе.

Метрика Пратта, рассмотренная в предыдущем разделе, была адаптирована для целей оценки качества сегментации. Одна из таких версий имеет вид [17]:

$$FOM_e = \begin{cases} \dfrac{1}{N_e} \displaystyle\sum_{i=1}^{N_e} \frac{1}{1+\gamma d_i^2}, & N_e > 0 \\ 1, & N_e = 0 \end{cases},$$

где N_e – число ошибочно классифицированных пикселей, d_i – расстояние i-го пикселя изображения до ближайшего пиксела, отнесённого к тому же классу на эталонном изображении, γ – масштабный множитель. При идеальной сегментации FOM_e=1.

Для оценки качества алгоритмов сегментации наряду со статистическими мерами и мерами оценки ло-

кализации используются и другие классы мер. Очевидно, что при хорошей сегментации эталонное и результирующее изображение должны иметь одинаковую степень фрагментации, т. е. количество сегментов на них должно совпадать (или почти совпадать). Для оценки степени фрагментации изображения в [17] также было предложено использовать следующую меру:

$$FRAG = \frac{1}{1+\left|\alpha(n_R - n_I)\right|^{\beta}},$$

где n_R – количество сегментов на результирующем изображении, n_I – количество сегментов на эталонном изображении, α, β – масштабные параметры (в [17] $\alpha = 0{,}16$, $\beta = 2$). Параметр α определяет вклад величины $(n_R - n_I)$ в значение *FRAG*, а параметр β определяет, насколько сильно штрафуются большие отклонения n_R от n_I по сравнению с малыми. Однако данная мера не учитывает информацию о значениях использовавшихся при сегментации характеристик изображения, полученных для классифицируемых пикселей. Поэтому в той же работе была предложена другая мера – *FOC* (figure of certainty), при вычислении которой такая информация, а именно интенсивность, используется:

$$FOC = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\left|\psi(f_i - \mu_j)\right|^{\delta}},$$

где N – количество пикселей в изображении, f_i – значение интенсивности пикселя i исходного изображения, μ_j – репрезентативное значение интенсивности j-го сегмента, к которому i-й пиксель был отнесён при сегментации, ψ, δ – масштабные параметры (их смысл тот же, что у α и β из предыдущей формулы). В простейшем случае, когда исходное изображение состоит из нескольких областей постоянной интенсивности, выражение $(f_i - \mu_j)$ представляет собой разность интенсивностей того сегмента, к которому i-й пиксель принадлежит на самом деле, и того, к которому он был отнесён в процессе сегментации. Данная мера допускает обобщение на случай, когда изображение содержит неоднородные области, и сегментация осуществляется на основе некоторого признака F (например, текстурной характеристики). Тогда f_i будет значением этого признака в пикселе i, а в качестве μ_j можно использовать среднее значение признака F для j-го сегмента.

Другая разновидность мер эффективности касается точности определения характеристик сегментов. Как известно, сегментация является начальным этапом анализа изображения. Её цель – представить изображение в виде, удобном для определения его количественных характеристик, используемых в дальнейшем для анализа этого изображения. С этой точки зрения, логично оценивать качество сегментации по тому, насколько точно (по сравнению с эталонным изображением) можно на сегментированном изображении определить такие характеристики. В работе [18] было предложено два критерия, основанных на таком подходе – *AUMA* (absolute ultimate measurement accuracy) и *RUMA* (relative ultimate measurement accuracy):

$$AUMA_f = \left|R_f - S_f\right|. \quad RUMA_f = \left(\left|R_f - S_f\right| \Big/ R_f\right)\times 100,$$

где R_f – значение признака f, полученное на эталонном изображении (геометрического, цвето-яркостного или текстурного), S_f – значение признака f, полученное на сегментированном изображении. По существу, мы имеем не *2*, а *2P* критерия оценки, где P – количество выделяемых признаков. Очевидно, что чем ближе значение *AUMA* и *RUMA* к нулю, тем выше оценка качества сегментации. Эти критерии можно использовать для оценки значимости того или иного признака для качества сегментации и, следовательно, для анализа изображения.

Приведённые выше количественные критерии оценки качества алгоритмов сегментации относятся к контролируемым эмпирическим методикам оценки и базируются на сравнении полученного сегментированного изображения с эталонной сегментации на соответствующем ground truth образе. Автоматические (неконтролируемые) методики оценки алгоритмов сегментации базируются не на сравнении с эталонной сегментацией, а на субъективных представлениях о том, какими свойствами должна обладать «хорошая сегментация». Критерии, используемые в этих методиках, и направлены на количественную оценку таких желаемых свойств. Отсутствие необходимости в эталонном изображении имеет свои потенциальные плюсы. Оно позволяет оценивать работу алгоритмов сегментации в режиме онлайн, что делает возможным, например, «на лету» настраивать параметры алгоритма в зависимости от оценки промежуточных результатов или определять, когда можно остановить итеративный процесс сегментации, достигнув некоего желаемого уровня качества.

Можно выделить три основные группы автоматических критериев, ориентированных на количественную оценку определённых свойств сегментированного изображения:

– однородность сегментов;

– контраст между соседними сегментами;

– форма сегментов.

На практике количественные меры эффективности, учитывающие один из перечисленных критериев, дают обычно менее адекватный результат, чем меры, использующие ground truth образы (что подтверждается визуальной оценкой качества сегментации). Кроме того, приведённые критерии носят субъективный характер и, следовательно, плохо формализуются. Таким образом, контролируемые методики оценки алгоритмов сегментации являются на данный момент более качественными. В последнее время появились комплексные меры эффективности, учитывающие более одного критерия, например, степень однородности сегментов, их количество (подробнее см. [19]).

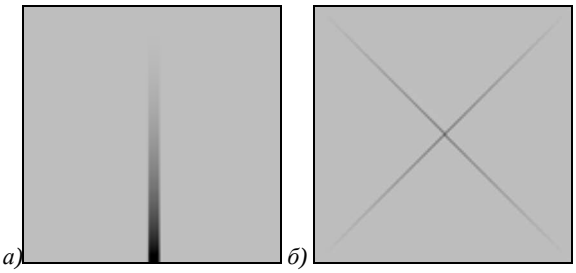
Следует отметить, что сколько-нибудь масштабных сравнительных исследований различных критериев оценки алгоритмов сегментации пока ещё не проводилось. В опубликованных работах или используется крайне ограниченный набор тестовых изобра-

жений, или исследуются критерии, относящиеся к одной группе и оценивающие одни и те же свойства сегментированного изображения. Это снова говорит об актуальности задачи совершенствования методологии сравнительного тестирования.

4. Некоторые аспекты методологии EDEM в системе PICASSO

В последние годы в НИИСИ РАН для сравнительного исследования алгоритмов обработки и анализа изображений разрабатывается программная среда PICASSO (PICture Algorithms Study SOftware). Цель данной деятельности – создание инструмента для разработки адаптивных систем анализа изображений для широкого спектра прикладных задач. Изначально среда PICASSO разрабатывалась для тестирования детекторов границ (этот её компонент в настоящее время наиболее развит). Её последующие версии также включили в себя тестирование алгоритмов реставрации изображений, анализа текстур, сегментации изображений. В настоящее время она включает в себя базу тестовых и эталонных изображений, базу текстур, редактор изображений, различные фильтры и генераторы шумов, шаблоны для заполнения фона и ряд других компонентов.

Идеологической основой системы PICASSO является методология EDEM (Empirical Discrepancy Evaluation Method), также находящаяся в процессе совершенствования. Учитывая, что тестируются алгоритмы, решающие разнообразные задачи обработки и анализа изображений, и теоретические аспекты данных областей пока недостаточно проработаны (например, не существует единого определения проблемы сегментации изображения, и в статьях можно найти не менее десятка различных определений, см. [20]), было решено не использовать аналитические методики оценки алгоритмов. В начальной версии системы PICASSO для тестирования методов выделения границ использовался набор специально разработанных двумерных искусственных изображений для моделирования различных ситуаций. Каждое из этих изображений моделирует ситуацию, в некотором смысле трудную для тестируемых алгоритмов, см. рис. 3*а-б*. Здесь трудность ситуации обусловлена наличием границы изменяющегося контраста.



а) *б)*

Рис. 3. Примеры изображений из тестового набора PICASSO: а) исчезающая полоса, б) затухающий узел

Процесс тестирования включал в себя:

– выбор алгоритма или группы похожих алгоритмов, которые предполагается тестировать;

– выбор из базы данных семейства искусственных тестовых изображений, моделирующих требуемые ситуации и соответствующих ground truth образов;

– оптимизацию параметров алгоритмов. Желательно ввести параметры так, чтобы один параметр оказывал влияние на одно и то же свойство всех тестируемых алгоритмов. В этом случае появляется возможность строить общие графики, на которых по горизонтальной оси откладывается этот параметр, а по вертикальной – отклики алгоритмов;

– выбор способов искажения тестовых изображений. Например, можно использовать разные методы зашумления или методы задания фона для изображения. Желательно, чтобы в этих способах были параметры, по которым можно строить графики: например, девиация шума;

– выбор метрики, дающей количественную оценку качества работы алгоритмов;

– статистическую обработку результатов.

Эта методика была использована для тестирования ряда детекторов границ на размытых и зашумлённых изображениях [21]. В данной работе тестировались алгоритмы Canny, Rothwell, Heitger, Black, Iverson и Smith, различные по своей сути, но решающие одну и ту же задачу. При тестировании всех перечисленных алгоритмов использовались их параметры по умолчанию. В качестве параметров искажения использовались различные величины дисперсии гауссовского шума и размера окна осреднения (параметр размытия). В качестве метрик эффективности использовались определённые выше чувствительность и специфичность. Уже в рамках такой методики был получен ряд содержательных результатов. Она дала принципиальную возможность сравнивать результаты тестирования в автоматическом режиме. Графическое представление результатов [21, 26] дало возможность качественно оценить поведение алгоритмов.

Кроме того, статистическая обработка количественных результатов тестирования (значений метрик) дало возможность выявить лидеров (алгоритмы Canny и Rothwell) и аутсайдеров (алгоритм Iverson) среди тестируемого набора. Вместе с тем данная методика имеет и ряд недостатков. Как было отмечено выше, использование только статистических мер (в данном случае чувствительности и специфичности) даёт мало информации о способности алгоритмов сохранять форму границ объектов. Использование в тестировании только изображений, содержащих трудные в распознавании ситуации, является разновидностью тестирования в наихудших случаях (worst case testing) и содержит недостатки, присущие данному подходу, который, в частности, не учитывает многообразия реальных ситуаций. Наконец, в процессе тестирования не было попыток отделить проблемы собственно алгоритмов от проблем в их программной реализации.

Указанные недостатки ранней версии EDEM были частично преодолены при дальнейшем развитии системы PICASSO. Здесь можно отметить работу [22], посвящённую исследованию детекторов границ при

аффинных преобразованиях: сдвигах, поворотах, сжатиях/растяжениях объектов на исходных изображениях. Для задачи распознавания объектов, размеры и ориентация которых заранее неизвестны, данное исследование актуально. Тестировался тот же набор алгоритмов, что и в [21]. Методологические изменения в сравнительном исследовании алгоритмов коснулись как тестовых изображений, так и используемых мер эффективности. Помимо тестовых изображений, моделирующих сложные для распознавания ситуации, использовались их упрощённые аналоги. В качестве примера на рис. 4*б* приведено упрощённое изображение, соответствующее образу Исчезающая полоса (рис. 3*а*). Здесь контраст границы постоянен.

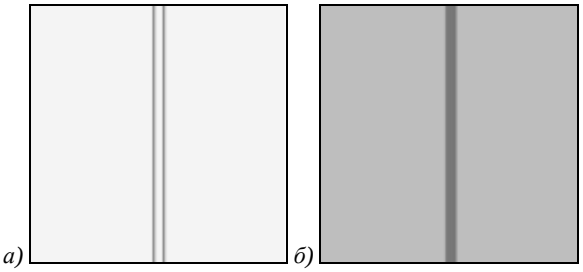


Рис. 4. Изображения, соответствующие изображению на рис. 3а (исчезающая полоса): а) ground truth образ, б) упрощённая версия

В качестве мер эффективности, наряду с чувствительностью и специфичностью, использовалась метрика Пратта (мера оценки локализации). Метрика Хаусдорфа использовалась в качестве вспомогательной. Результаты тестирования на упрощённых изображениях показали существенно худшее поведение двух из рассматриваемых детекторов. Остальные четыре детектора показали приемлемые результаты (что подтвердила и визуальная оценка). При этом оказалось невозможно выявить лидера среди них, опираясь на значения данных мер эффективности. Также было замечено, что в некоторых случаях выбор между результатами работы двух детекторов на одном тестовом образе осуществлялся на основании значений метрики Хаусдорфа, в то время как значения других метрик практически совпадали (и визуально результаты работы были неразличимы). Здесь метрика Хаусдорфа сыграла роль своего рода «увеличительного стекла». Таким образом, использование метрики Хаусдорфа в сочетании с другими метриками оказывалось полезным (это, в частности, опровергает вывод [14] о практической непригодности метрики Хаусдорфа для тестирования детекторов границ).

Тестирование всех алгоритмов на исходных изображениях, содержащих границы переменного контраста, дало в целом неудовлетворительные результаты. Визуальная оценка это подтвердила (один такой пример изображён на рис. 5).

Что касается проблемы отделения тестирования собственно алгоритмов от тестирования их программных реализаций, был предложен ряд простых тестов для выявления ошибок последних. Например, тестировалось поведение алгоритмов при повороте объекта на изображении на 180 градусов (результаты тестирования обработанных исходного и повёрнутого изображений должны быть близкими к идентичным). Также в работе был получен ряд плохих результатов тестирования алгоритма Canny при поворотах объектов на упрощённых изображениях. Учитывая известность данного алгоритма (фактически здесь используется опыт прошлого тестирования), было сделано естественное предположение, что причина неадекватного поведения – его программная реализация. После её замены на версию алгоритма из MATLAB, результаты оказались одними из лучших.

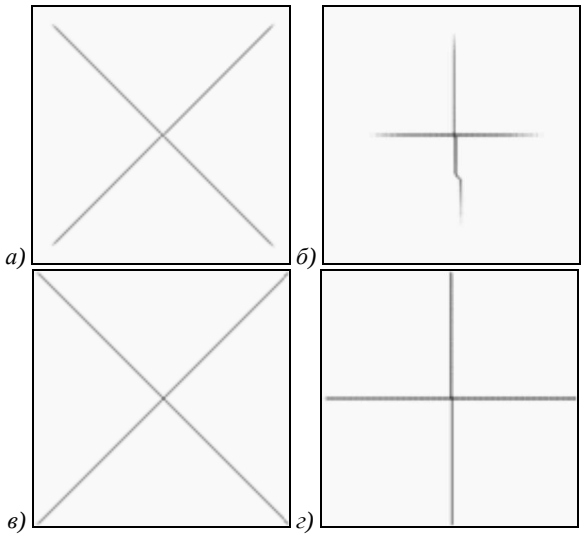


Рис. 5. Результаты обработки затухающего узла (рис. 3б) алгоритмом Smith: а) исходное изображение, б) его поворот на 45 градусов; в) и г) результаты обработки упрощенной версии изображения

Таким образом, методологические изменения, выраженные в сочетании исходных и упрощённых изображений и различных мер эффективности, позволили улучшить качество оценки алгоритмов и дополнительно получить новые практически важные результаты. Например, если местоположение и относительные размеры объекта на изображении заранее неизвестны, но известно, что контраст границы объекта меняется незначительно, в качестве детектора границ объекта допустимо использовать четыре из шести протестированных алгоритма с параметрами, заданными по умолчанию. Если же контраст границы объекта претерпевает существенные изменения, то их использование не рекомендуется (как минимум потребуется процедура настройки параметров алгоритма).

Выше отмечалось, что полутоновые изображения являются нечёткими по своей природе из-за неопределённости, существующей в локализации положения границы, отделяющей объект от фона. Также при обработке данных дистанционного зондирования из-за несовершенства съёмочной аппаратуры бывает, например, сложно отнести тот или иной пиксель на изображении к области городской застройки, пашни или лесов. Эта неопределённость наводит на мысль о возможности использования элементов теории нечётких множеств в задачах анализа изображений, в част-

ности, в задачах выделения границ и сегментации. В последнее время для решения данных задач, а также для задач уточнения границ, реставрации изображений, анализа текстур был разработан ряд алгоритмов, использующих элементы нечёткой логики. Для тестирования этих алгоритмов потребовалось внести в методологию EDEM соответствующие дополнения [13], [23]. Они позволили усовершенствовать тестирование и традиционных, «чётких» алгоритмов. Причём в рамках подхода, предложенного в данных работах, можно одновременно оценивать «чёткие» и «нечёткие» алгоритмы и сравнивать между собой результаты их оценки.

Следует отметить, что методика сравнительного исследования «чётких» и «нечётких» алгоритмов анализа изображений не получила к настоящему моменту достаточного развития. В ряде работ, посвящённых анализу и классификации аэрокосмических снимков, было проведено обобщение нескольких известных статистических мер оценки качества (подобных тем, что применяются для оценки качества сегментации, см. выше) на случай нечётких множеств, позволяющее в том числе сравнивать между собой чёткие и нечёткие множества. Соответствующие меры были названы нечёткими мерами сходства (fuzzy similarity measures). Там же рассматривалась и концепция нечёткого ground truth образа. Результаты данных работ предназначались для определения на снимках областей лесов, воды и т. п., но они также могут быть применены для исследования эффективности различных детекторов границ. Это наблюдение нашло своё отражение в работе [23].

Что касается нечетких ground truth образов, применяемых для тестирования детекторов границ, они отождествляются с функцией принадлежности пикселей изображения множеству границ, принимающей значения от 0 до 1. При этом обычные ground truth образы, содержащие эталонную карту границ, отождествляются с её характеристической функцией (принимающей значение 1 для граничных пикселей и 0 для пикселей, не являющихся граничными). Нечёткие ground truth образы применяемых для тестирования алгоритмов сегментации отождествляются с набором функций принадлежности сегментам.

В рамках совершенствования методологии EDEM в [23] было предложено использовать различные нечёткие ground truth образы, соответствующие одному и тому же тестовому изображению. Предполагалось, что они позволят лучше протестировать те или иные свойства исследуемого детектора границ. В результате, например, оказалось, что одни такие образы лучше использовать для проверки способности тестируемого детектора выделять слабые края, в то время как другие более приспособлены для проверки способности детектора выделять непрерывные границы (функции принадлежности, соответствующие этим образам, чувствительны к разрывам (дыркам) на карте границ).

Другим потенциально важным приложением нечётких ground truth образов, отмеченным в [23], является возможность их использования в проверке свойства детектора границ выделять граничные точки, существенные для определения ограничиваемого объекта (т.н. image feature points). Например, для прямоугольника таковыми являются угловые точки. Задавая для таких точек более высокие значения функции принадлежности множеству границ в сравнении с остальными граничными точками ground truth образа, можно добиться высокой чувствительности нечётких мер сходства к выделению этих существенных точек. Отметим, что и для тестирования алгоритмов сегментации использование нечётких ground truth образов представляется перспективным. В настоящее время нами разрабатывается методика построения нечётких ground truth образов и использования различных нечётких мер сходства для практического тестирования детекторов границ и алгоритмов сегментации изображений.

Заключение и выводы

В последние два десятилетия проблеме сравнительного исследования эффективности алгоритмов анализа изображений стало уделяться всё больше внимания в научной литературе. Поскольку общей теории обработки и анализа изображений все ещё не существует, то аналитические методики оценки алгоритмов малоприменимы и имеют дело лишь с задачами специального вида. Основное внимание в настоящий момент уделяется развитию эмпирических методик, как использующих для оценки результатов работы алгоритмов эталонные ground truth изображения (контролируемые методики), так и оценивающих качество работы алгоритма непосредственно, без сравнения с эталоном (автоматические методики). Первые из этих методик дают более точную оценку, а вторые могут использоваться в системах реального времени для оценки алгоритма в онлайн-режиме.

Характеризуя разрабатываемую нами методологию эмпирической оценки алгоритмов анализа изображений EDEM, можно выделить следующие её основные характеристики:

– использование тестовых изображений, моделирующих трудные для алгоритма ситуации;

– сочетание сложных тестовых изображений с их упрощёнными версиями;

– использования метрик разных классов для количественной оценки качества алгоритмов (например, использование статистических мер и мер оценки локализации при оценке детекторов границ);

– организация процесса сравнительного тестирования, дающая возможность качественного анализа его результатов (например, построения графиков для сравнительного их анализа);

– использование для тестирования элементов теории нечётких множеств. Разработка нечётких ground truth образов. Использование нескольких таких образов, соответствующих одному тестовому изображению для анализа различных свойств тестируемого алгоритма. Адаптация имеющихся нечётких мер сход-

ства (fuzzy similarity measures) к задаче оценки алгоритмов анализа изображений.

Данная методология прошла апробацию при решении практических задач (последняя из которых – разработка технологии сегментации изображений клеток крови, [24]–[25]). Тем не менее ряд вопросов, относящихся как к общей методологии сравнительного исследования алгоритмов, так и к методологии EDEM, остаётся открытым (подробнее см. [26]). Например, существенной трудностью, возникающей при оценке алгоритмов, является выбор подходящей метрики, дающей объективную меру их эффективности. На практике метрика даёт количественную оценку, призванную характеризовать те или иные аспекты эффективности алгоритма. При этом ни одна метрика не даёт исчерпывающего ответа. Таким образом, разработка мер эффективности и сочетание различных мер в рамках одной методологии тестирования является одной из самых актуальных задач сравнительного исследования алгоритмов обработки изображений. В особенности это касается вопроса сочетания различных однотипных мер (например, мер оценки локализации). Использование нескольких мер одного класса неизбежно приводит к вопросу об их ранжировании (какой из них можно доверять больше). В рамках методологии EDEM получен положительный результат использования метрики Хаусдорфа в качестве вспомогательной (второй) меры оценки локализации при оценке детекторов границ и некоторых алгоритмов сегментации [20], [22].

Что касается методологии применения тестовых изображений и соответствующих им ground truth образов, одним из основных открытых вопросов здесь является вопрос о полноте данного набора в задачах тестирования. Наша эмпирическая методология подразумевает выбор тестовых изображений исходя из специфики конкретной задачи. Например, при тестировании детекторов границ должна учитываться плотность границ на тестовых изображениях, соответствующая плотности границ на реальных изображениях, для которых тестируемые детекторы предназначаются. Соответственно, в рамках программной системы PICASSO совершенствуется технология создания тестовых изображений с учётом приведённых выше факторов. Также выше говорилось об использовании в рамках методологии EDEM нескольких нечётких ground truth образов, соответствующих одному тестовому изображению. На данный момент остаётся открытым вопрос создания методики построения таких образов и использования различных нечётких мер сходства для практического тестирования. Наконец, опыт наших практических исследований сравнительного тестирования алгоритмов показывает, что в ряде случаев невозможно выделить абсолютного лидера по результатам, полученным на всей совокупности тестируемых изображений. При этом есть несколько алгоритмов, показывающих близкие результаты. В такой ситуации при выборе наиболее пригодного в практическом использовании алгоритма на первый план выходят такие его свойства, как сложность в использовании, ресурсоёмкость, возможность распараллеливания и другие подобные характеристики, являющиеся элементами аналитической методологии оценки алгоритма. Мы планируем разработать методику оценки этих характеристик в дальнейших версиях нашей системы.

Благодарности

Работа выполнена при поддержке гранта РФФИ № 14-07-00502.

Литература

1. **Kirsch, R.A.** Experiments in processing pictorial information with a digital computer / R.A. Kirsch, L. Kahn, C. Ray, G.H. Urban // Proceedings of the Eastern Joint Computer conference. – 1957. – P. 221-229.
2. **Zhang, Y.J.** Evaluation and comparison of different segmentation algorithms / Y.J. Zhang // Pattern Recognition Letters. – 1997. – Vol. 18(10). – P. 963-974. – ISSN 0167-8655.
3. **Heath, M.D.** Robust visual method for assessing the relative performance of edge detection algorithms /M.D. Heath, S. Sarkar, T. Sanocki, K. Bowyer // IEEE Transactions on Pattern Analysis and Machine Intelligence – 1997. – Vol. 19(12). – P. 1338-1359. – ISSN 0167-8655.
4. **Wirth, M.A.** Performance evaluation of image processing algorithms in CADe / M.A. Wirth // Technology in Cancer Research and Treatment. – 2005. – Vol. 4(2). – P. 159-172. – ISSN 1533-0346.
5. **Zhang, H.** Image segmentation evaluation: A survey of unsupervised methods / H. Zhang, J.E. Fritts, S.A. Goldman // Computer Vision and Image Understanding. – 2008. – Vol. 110(2). – P. 260-280. – ISSN 1077-3142.
6. **Shin, M.C.** Comparison of edge detector performance through use in an object recognition task / M.C. Shin, D. Goldgof, K. Bowyer // Computer Vision and Image Understanding. – 2001. – Vol. 84(1). – P. 160-178. – ISSN 1077-3142.
7. **Cardoso, J.S.** Toward a Generic Evaluation of Image Segmentation / J.S. Cardoso, L. Corte-Real // IEEE Transactions on Image Processing. – 2005. – Vol. 14(11). – P. 1773-1782. – ISSN 1057-7149.
8. **Thomas, G.A.** 3d image sequence acquisition for tv and film production / G.A. Thomas, O. Grau // Proceedings of 1st International Symposium on 3D Data Processing, Visualisation and Transmission. – 2002. – P. 320-326.
9. **Zhang, Y.J.** A survey on evaluation methods for image segmentation / Y.J. Zhang // Pattern Recognition. – 1996. – Vol. 29(8). – P. 1335-1346. – ISSN 0031-3203.
10. **Zhang, Y.J.** Image segmentation evaluation in this century / Y.J. Zhang // Encyclopedia of Information Science and Technology. Editor M. Khosrow-Pour. 2nd edition. – IGI Global, 2009. – P. 1812-1817.
11. **Haralick, R.M.** Image segmentation techniques / R.M. Haralick, L.G. Shapiro // Computer Vision, Graphics, and Image Processing. – 1985. – Vol. 29(1). – P. 100-132. – ISSN 0734-189X.
12. **Canny, J.** A computational approach to edge detection / J. Canny // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1986. – Vol. 8(6). – P. 679-698. – ISSN 0167-8655.
13. **Грибков, И.В.** Некоторые вопросы количественной оценки производительности детекторов границ / И.В. Грибков, А.В. Захаров, П.П. Кольцов, Н.В. Котович, А.А. Кравченко, А.С. Куцаев, А.С. Осипов //

Программные продукты и системы. – 2011. – № 4. – С. 13-20.
14. **Baddeley, A.J.** Errors in binary images and L^p version of the Hausdorff Metric / A.J. Baddeley // Nieuw Archief voor Wiskunde. – 1992 – Vol. 10. – P. 157-183. – ISSN 0028-9825.
15. **Yasnoff, W.A.** Error measures for scene segmentation / W.A. Yasnoff, J.K. Mui, J.W. Bacus // Pattern Recognition. – 1977, – Vol. 9(4). – P. 217-231. – ISSN 0031-3203.
16. **Van Droogenbroeck, M.** Design of Statistical Measures for the Assessment of Image Segmentation Schemes / M. Van Droogenbroeck, O. Barnich // Proceedings of 11th International Conference on Computer Analysis of Images and Patterns (CAIP2005), Lecture Notes in Computer Science. – 2005, – Vol. 3691. – P. 280-287.
17. **Strasters, K.S.** Three-dimensional image segmentation using a split, merge and group approach / K.C. Strasters, J.J. Gerbrands // Pattern Recognition Letters. – 1991. – Vol. 12(5). – P. 307-325. – ISSN 0167-8655.
18. **Zhang, Y.J.** Objective and quantitative segmentation evaluation and comparison / Y.J. Zhang, J.J. Gerbrands // Signal Processing. – 1994. – Vol. 39(1-2). – P. 43-54. – ISSN 0165-1684.
19. **Захаров, А.В.** Критерии оценки качества сегментации изображений. / А.В. Захаров, П.П. Кольцов, Н.В. Котович, А.А. Кравченко, А.С. Куцаев, А.С. Осипов // Труды НИИСИ РАН. – 2012 – Т. 2, № 2 – С. 87-99.
20. **Грибков, И.В.** Тестирование методов сегментации изображений в системе PICASSO / И.В. Грибков, А.В. Захаров, П.П. Кольцов, Н.В. Котович, А.А. Кравченко, А.С. Куцаев, А.С. Осипов – М.: НИИСИ РАН, 2007.
21. **Gribkov, I.V.** PICASSO – A System for Evaluating Edge Detection Algorithms / I.V. Gribkov, P.P. Koltsov, N.V. Kotovich, A.A. Kravchenko, A.S. Kutsaev, V.K. Nikolaev, A.V. Zakharov // Pattern Recognition and Image Analysis. – 2003. – Vol. 13(4). – P. 617-622.
22. **Gribkov, I.V.** Edge Detection under Affine Transformations: Comparative Study by PICASSO 2 System / I.V. Gribkov, P.P. Koltsov, N.V. Kotovich, A.A. Kravchenko, A.S. Kutsaev, A.S.Osipov, A.V. Zakharov // WSEAS Transactions on Signal Processing. – 2006. – Vol. 2(9). – P. 1215-1221.
23. **Osipov, A.** A fuzzy approach to performance evaluation of edge detectors /A. Osipov // Lecture Notes in Signal Science, Internet and Education. – WSEAS Press, 2007. – P. 94-99.
24. **Koltsov, P.** On one approach to blood cell image segmentation / P. Koltsov, N. Kotovich, A. Kravchenko, A. Koutsaev, A. Kuznetsov, A. Osipov, E. Sukhenko, A. Zakharov // The 11th International Conference "Pattern Recognition and Image Analysis" (PRIA-11-2013), Conference Proceedings. – 2013. – Vol. 2. – P. 615-618.
25. **Беляков, В.К.** Об одной методике классификации клеток крови и ее программной реализации / В.К. Беляков, Е.П. Сухенко, А.В. Захаров, П.П. Кольцов, Н.В. Котович, А.А. Кравченко, А.С. Куцаев, А.С. Осипов, А.Б. Кузнецов // Программные продукты и системы. – 2014. – № 4. – С. 46-56.
26. **Захаров, А.В.** Прямая оценка качества программных продуктов. Критерии и тестовые материалы / А.В. Захаров, П.П. Кольцов, Н.В. Котович, А.А. Кравченко, А.С. Куцаев, А.С. Осипов // Программные продукты, системы и алгоритмы. – 2014. – № 3. – С. 1-8.

References

[1] Kirsch RA, Kahn L, Ray C, Urban GH. Experiments in processing pictorial information with a digital computer. Proceedings of the Easterm Joint Computer conference 1957; 221-9.
[2] Zhang YJ. Evaluation and comparison of different segmentation algorithms. Pattern Recognition Letters 1997; 18(10): 963-74.
[3] Heath MD, Sarkar S, Sanocki T, Bowyer K. Robust visual method for assessing the relative performance of edge detection algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 1997; 19(12): 1338-59.
[4] Wirth MA. Performance evaluation of image processing algorithms in CADe. Technology in Cancer Research and Treatment 2005; 4(2): 159-72.
[5] Zhang H, Fritts JE, Goldman SA. Image segmentation evaluation: A survey of unsupervised methods. Computer Vision and Image Understanding 2008; 110(2): 260-80.
[6] Shin MC, Goldgof D, Bowyer K. Comparison of edge detector performance through use in an object recognition task. Computer Vision and Image Understanding 2001; 84(1): 160-78.
[7] Cardoso JS, Corte-Real L. Toward a Generic Evaluation of Image Segmentation. IEEE Transactions on Image Processing 2005; 14(11): 1773-82.
[8] Thomas GA, Grau O. 3d image sequence acquisition for tv and film production. Proceedings of 1st International Symposium on 3D Data Processing, Visualisation and Transmission 2002; 320-6.
[9] Zhang YJ. A survey on evaluation methods for image segmentation. Pattern Recognition 1996; 29(8): 1335-46.
[10] Zhang YJ. Image segmentation evaluation in this century. Encyclopedia of Information Science and Technology. 2nd edition. IGI Global; 2009: 1812-17.
[11] Haralick RM, Shapiro LG. Image segmentation techniques Computer Vision, Graphics, and Image Processing 1985; 29(1): 100-32.
[12] Canny J. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 1986; 8(6): 679-98.
[13] Gribkov IV, Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Osipov AS, Zakharov AV**.** On some issues of the quantitative performance evaluation of edge detectors [In Russian]. Programmnye Produkty I Sistemy 2011; 2: 13-9.
[14] Baddeley AJ. Errors in binary images and L^p version of the Hausdorff Metric. Nieuw Archief voor Wiskunde 1992; 10: 157-83.
[15] Yasnoff WA, Mui JK, Bacus JW. Error measures for scene segmentation. Pattern Recognition 1977; 9(4): 217-31.
[16] Van Droogenbroeck M, Barnich O**.** Design of Statistical Measures for the Assessment of Image Segmentation Schemes. Proceedings of 11th International Conference on Computer Analysis of Images and Patterns (CAIP2005), Lecture Notes in Computer Science 2005; 3691: 280-7.
[17] Strasters KS, Gerbrands JJ**.** Three-dimensional image segmentation using a split, merge and group approach. Pattern Recognition Letters 1991; 12(5): 307-25.
[18] Zhang YJ, Gerbrands JJ**.** Objective and quantitative segmentation evaluation and comparison. Signal Processing 1994; 39(1-2): 43-54.
[19] Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Osipov AS, Zakharov AV**.** Criteria for evaluating image segmentation [In Russian]. Trudy NIISI RAN 2012; 2(2): 87-99.
[20] Gribkov IV, Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Osipov AS, Zakharov AV**.** Testing of image segmentation methods in PICASSO system [In Russian]. Moscow: NIISI RAN; 2007.

[21] Gribkov IV, Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Nikolaev VK, Zakharov AV. PICASSO – A System for Evaluating Edge Detection Algorithms. Pattern Recognition and Image Analysis 2003; 13(4): 617-22.
[22] Gribkov IV, Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Osipov AS, Zakharov AV. Edge Detection under Affine Transformations: Comparative Study by PICASSO 2 System. WSEAS Transactions on Signal Processing 2006; 2(9): 1215-21.
[23] Osipov A. A fuzzy approach to performance evaluation of edge detectors. Lecture Notes in Signal Science, Internet and Education. WSEAS Press; 2007: 94-9.
[24] Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Kuznetsov AB, Osipov AS, Sukhenko EP, Zakharov AV. On one approach to blood cell image segmentation. The 11th International Conference "Pattern Recognition and Image Analysis" (PRIA-11-2013), Conference Proceedings 2013; 2: 615-18.
[25] Belyakov VK, Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Kuznetsov AB, Osipov AS, Sukhenko EP, Zakharov AV. On one method of blood cell classification and its software implementation [In Russian] Programmnye Produkty I Sistemy 2014; 2: 46-56.
[26] Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Osipov AS, Zakharov AV. Direct assessment of software quality. Criteria and materials for testing [In Russian]. Programmnye Produkty, Sistemy I Algoritmy 2014; 3: P. 1-8.

ON THE QUANTITATIVE PERFORMANCE EVALUATION OF IMAGE ANALYSIS ALGORITHMS

P.P. Koltsov[1], A.S. Osipov [1], A.S. Kutsaev [1], A.A. Kravchenko [1], N.V. Kotovich[1], A.V. Zakharov[1]
[1] Scientific-Research Institute for System Analysis, Russian Academy of Sciences

Abstract

The paper contains a brief review of main approaches to the comparative performance evaluation of image analysis algorithms. Some empirical methods used for the comparative evaluation of edge detectors and image segmentation algorithms are considered and quantitative criteria employed in these methods are studied. Problems associated with the use of these criteria are described. Finally, using the edge detector evaluation as an example, we propose an empirical method, called EDEM, which is implemented using our proprietary software system PICASSO.

Keywords: comparative study, image analysis, edge detectors, image segmentation, performance measures, ground truth image, fuzzy sets.

Citation: Koltsov PP, Osipov AS, Kutsaev AS, Kravchenko AA, Kotovich NV, Zakharov AV. On the quantitative performance evaluation of image analysis algorithms. Computer Optics 2015; 39(4): 542-56. DOI: 10.18287/0134-2452-2015-39-4-542-556.

Сведения об авторах

Кольцов Петр Петрович, 1946 года рождения, в 1971 году окончил Московский физико-технический институт по специальности «Динамика полёта и управление». Кандидат физико-математических наук (1975), доктор технических наук (2012). Автор более 80 статей и 1 монографии. Научные интересы: распознавание образов, обработка изображений, математическое моделирование.

E-mail: *kppkpp@mail.ru* .

Piotr Petrovich Koltsov (b. 1946) graduated from Moscow Institute of Physics and Technology in 1971, majoring in Flight Dynamics and Control. He received his Candidate of Science degree (Physics and Mathematics) in 1975 and Doctor of Technical Sciences degree in 2012. Author of more than 80 papers and 1 monograph. Research interests are pattern recognition, image processing and mathematical modeling.

Осипов Андрей Сергеевич, 1968 года рождения, в 1990 году окончил Московский государственный университет по специальности «Математика». Кандидат физико-математических наук (1995), работает старшим научным сотрудником в Научно-исследовательском институте системных исследований РАН. Научные интересы: теория операторов, обратные задачи, обработка изображений.

E-mail: *osipa@niisi.ras.ru* .

Andrey Sergeevich Osipov (b. 1968) graduated from Moscow State University in 1990, majoring in Mathematics. He received his Candidate of Science degree (Physics and Mathematics) in 1995. Currently he works as senior researcher at Institute for System Studies of the Russian Academy of Sciences. Research interests are operator theory, inverse problems and image processing.

Куцаев Александр Сергеевич, 1951 года рождения, в 1973 году окончил Московский государственный университет по специальности «Механика». Кандидат физико-математических наук (1982), работает старшим научным сотрудником в Научно-исследовательском институте системных исследований РАН. Научные интересы: распознавание образов, обработка изображений, программирование.

E-mail: *koutsaev@niisi.msk.ru* .

Aleksandr Sergeevich Koutsaev (b. 1951) graduated from Moscow State University in 1973, majoring in Mechanics. He received his Candidate of Science degree (Physics and Mathematics) in 1982. Currently he works as senior re-

searcher at Institute for System Studies of the Russian Academy of Sciences. Research interests are pattern recognition, image processing, and programming.

Кравченко Александр Анатольевич, 1958 года рождения, в 1980 окончил Московский государственный университет по специальности «Математика». Кандидат физико-математических наук (1984), работает заведующим отделом в Научно-исследовательском институте системных исследований РАН. Научные интересы: компьютерная графика, распознавание образов, обработка изображений.

E-mail: *alexk@niisi.msk.ru* .

Aleksandr Anatolevich Kravchenko (b. 1958) graduated from Moscow State University in 1980, majoring in Mathematics. He received his Candidate of Science degree (Physics and Mathematics) in 1984. Currently he works as head of department at Institute for System Studies of the Russian Academy of Sciences. Research interests are computer graphics, pattern recognition and image processing.

Котович Николай Владимирович, 1958 года рождения, в 1980 окончил Московский государственный университет по специальности «Математика». Работает старшим научным сотрудником в Научно-исследовательском институте системных исследований РАН. Научные интересы: распознавание образов, обработка изображений.

E-mail: *kotovich@niisi.msk.ru* .

Nikolay Vladimirovich Kotovich (b. 1958) graduated from Moscow State University in 1980, majoring in Mathematics. Currently he works as Senior Researcher at Institute for System Studies of the Russian Academy of Sciences. Research interests are pattern recognition and image processing.

Захаров Алексей Викторович, 1960 года рождения, в 1983 году окончил Московский государственный университет по специальности «Астрономия». Работает старшим научным сотрудником в Научно-исследовательском институте системных исследований РАН. Область научных интересов: обработка изображений, распознавание образов.

E-mail: *avz@compot.ru* .

Aleksei Viktorovich Zakharov (b. 1960) graduated from Moscow State University in 1983, majoring in Astronomy. Currently he works as senior researcher at Institute for System Studies of the Russian Academy of Sciences. Research interests are pattern recognition and image processing.

Поступила в редакцию 20 апреля 2015 г.
Окончательный вариант – 22 июля 2015 г.