УДК 004.932

ПРЯМАЯ ОЦЕНКА КАЧЕСТВА ПРОГРАММНЫХ ПРОДУКТОВ. КРИТЕРИИ И ТЕСТОВЫЕ МАТЕРИАЛЫ

Работа выполнена при поддержке РФФИ и РЖД,
проекты №№ 13-07-13128 офи_м_РЖД и 14-07-00502

А.В. Захаров, ст. научный сотрудник; П.П. Кольцов, д.т.н., доцент;
Н.В. Котович, ст. научный сотрудник; А.А. Кравченко, к.ф.-м.н., зав. сектором;
А.С. Куцаев, к.ф.-м.н., ст. научный сотрудник;
А.С. Осипов, к.ф.-м.н., ст. научный сотрудник
(НИИСИ РАН, Нахимовский просп., 36, корп. 1, г. Москва, 117218, Россия, zaharov@niisi.msk.ru, kppkpp@mail.ru, kotovich@niisi.msk.ru, alexk@genebee.msu.su, koutsaev@niisi.msk.ru, osipa68@yahoo.com)

Аннотация. Статья посвящена сравнительному исследованию некоторых методов обработки изображений, предназначенных для решения задач дистанционного мониторинга с целью обеспечения железнодорожной безопасности. Приводится информация об основных используемых в настоящее время методиках оценки программных продуктов в области обработки и анализа изображений. Основное внимание при этом уделено методам количественной оценки результатов обработки изображений, использующим ground truth-изображения и меры различий. Дано описание разработанной в НИИСИ РАН методики EDEM сравнительного исследования алгоритмов обработки изображений. В качестве примера использования методики EDEM рассмотрена проблема оценки качества программных продуктов, предназначенных для решения задачи сегментации изображений, поступающих с борта дистанционно-пилотируемого летательного аппарата. Цель данной сегментации заключается в выделении областей на изображении, существенных для принятия решений (например, относящихся к обеспечению безопасности) оператором дистанционно-пилотируемого летательного аппарата. Здесь подробно рассмотрены различные количественные критерии оценки, используемые в рамках данной методики. Сформулированы общие требования к тестовым материалам и ground truth-эталонам, выработанные при разработке методики EDEM. Приведены результаты применения данной методики, реализованной в программной среде PICASSO, к оценке программ сегментации изображений. В качестве другого возможного приложения методики EDEM к оценке программных продуктов в области дистанционного мониторинга рассмотрена проблема выделения на изображении важных (в контексте решаемой внешней задачи) элементов. Показано, что для оценки программных продуктов, предназначенных для решения данной проблемы, использование элементов теории нечетких множеств является перспективным.

Ключевые слова: *дистанционный мониторинг, сравнительное исследование, методы сегментации, детекторы границ, ground truth.*

Утверждение о том, что решение производственных задач, базирующееся на использовании программно-алгоритмических моделей, является надежным способом повышения качества и эффективности производства, давно является очевидным фактом. Формализацией и построением решений таких задач заняты многочисленные исследовательские коллективы. Однако сложность этих задач такова, что даже их формализация, необходимая для дальнейшего поиска обоснованного решения, вызывает затруднения, приводящие к многочисленным вариантам [1]. В свою очередь, собственно решение этих задач с использованием современных средств вычислительной техники приводит к возникновению различных программно-алгоритмических вариантов таких решений даже в рамках одной фиксированной формализации.

В связи с этим возникает новая проблема – сравнительная оценка качества решений некоторой фиксированной задачи различными программно-алгоритмическими методами, реализующими ту или иную формализацию этой задачи. Очевидно, что, только получив объективную оценку качества различных решений, можно определить, какое среди них наиболее эффективно с точки зрения производства. Именно эта проблема и пути ее решения будут далее рассмотрены на примере задач управления инфраструктурой железнодорожного транспорта, относящихся к сфере обеспечения железнодорожной безопасности.

В качестве таких задач, на которых были опробованы элементы разрабатываемой методики сравнительной оценки результатов работы программных средств EDEM, взяты задачи, относящиеся к дистанционному мониторингу в режиме реального времени. В последнее время для решения таких задач все активнее применяются *дистанционно-пилотируемые летательные аппараты* (ДПЛА). Так, компанией

Deutsche Bahn ДПЛА используется для борьбы с железнодорожными вандалами и рисовальщиками граффити, ежегодный ущерб от действий которых в Германии составляет примерно 10 млн. долларов [2]. В северо-восточной Индии, где в 2013 году произошло более 20 столкновений поездов со слонами, для мониторинга передвижения этих животных также используются ДПЛА [3].

Использование ДПЛА для решения задач дистанционного мониторинга подразумевает, что оперативное решение принимается оператором ДПЛА на основе визуальной информации, получаемой с борта. Для повышения качества таких решений обычно используется компьютерный анализ изображений, позволяющий в автоматическом режиме выделять на поступающих изображениях области, наиболее важные для принятия оперативных решений [4]. Надежность существующих в настоящее время компьютерных программ видеоконтроля недостаточно высока, поэтому проблема сравнительной оценки этих программ с целью выявления наиболее эффективных стоит весьма остро. Выбор вышеуказанных классов задач управления инфраструктурой железнодорожного транспорта был определен как их высокой приоритетностью в обеспечении комплексной многоуровневой безопасности движения [5], так и большим числом опубликованных вариантов решения таких задач. В ситуации, когда количество предложенных решений той или иной приоритетной задачи слишком велико, необходимость объективной оценки качества программных продуктов становится весьма актуальной. В работе выполнен сравнительный анализ компьютерных программ обработки изображений на основе подходов, используемых в разрабатываемой методике, элементы которой проиллюстрированы примерами.

Методики оценки программных продуктов

В настоящее время не существует единой методики оценки качества работы различных компьютерных программ, решающих некоторую содержательную задачу. Основные отличия методик, применяемых в сравнительных исследованиях различных программ, решающих задачу из рассматриваемых классов, заключаются в используемом типе критерия оценки качества (количественный или качественный, использующий эталонное решение или нет) и в используемом типе эталонов (реальные или синтезированные), их параметрах, количестве, источниках (оригинальные или общедоступные) и т.п.

Необходимо также отметить, что различные методики могут по-разному определять процедуру выбора оптимальных значений для управляющих параметров оцениваемой программы.

К настоящему времени предпринято несколько попыток классифицировать эти методики. В работе была использована достаточно популярная среди исследователей классификация методик сравнительного исследования алгоритмов, предложенная в [6], согласно которой методики оценки делятся на субъективные и объективные. Первые из них ориентированы на получение оценок качества на основе мнения экспертов и в данной работе не используются. Вторые подразделяются на системные, дающие оценку работы программы по результатам работы некоторой системы, в которую она входит как компонент, и прямые, имеющие дело непосредственно с исследуемой программой. В данной работе рассматриваются прямые методики, ориентированные на получение оценки качества работы программной реализации алгоритма, решающего конкретную задачу из области обработки и анализа изображений. Развитие таких методик оценки позволяет разработчикам сложных программных систем понимания изображений, использующих данные алгоритмы, оценивать промежуточные результаты. Это дает содержательную информацию о работе системы в целом и возможных путях ее совершенствования. Кроме того, развитие концепций, используемых в прямых объективных методиках (прежде всего концепций эталонов и метрик) полезно и для системных методик оценки. Очевидно, что в данной работе, ориентированной на получение объективной оценки качества работы компьютерной программы, используемая методика должна быть объективной и прямой. Среди таких методик различают аналитические и эмпирические [6]. Аналитические методики рассматривают алгоритм независимо от его выхода [7]. Изучаются такие свойства алгоритма, как стратегия реализации главной цели, сложность, возможность распараллеливания, ресурсоемкость и т.п. Эти свойства не имеют прямого отношения к качеству работы алгоритма. Эмпирические методики, напротив, оценивают не сам алгоритм, а результаты его работы на некотором наборе эталонов. Такие методики с помощью вариации эталонов позволяют оценить качество работы компьютерных программ на широком спектре внешних условий, учесть особенности практического применения программ, включая границы применимости. Эти обстоятельства определили выбор пути в разработке методики для поддержки объективного выбора программных и технических решений через построение объективной прямой эмпирической методики оценки качества программных продуктов. Более точно – разрабатываемая методика будет строить оценку качества на основе количественной меры различия результата работы программы на некотором наборе эталонов, для которых точное решение, так называемое ground truth, известно априори. Такой подход к оценке качества программных продуктов в англоязычной литературе обычно называется discrepancy method [8], а для собственно критерия используются термины evaluation criterion, performance criterion, performance metric, performance measure, performance index. Эти обстоя-

тельства определили выбор названия методики, применение элементов которой рассмотрено в работе: EDEM (Empirical Discrepancy Evaluation Method).

Критерии оценки качества результатов работы программ

Важным элементом методики EDEM является выбор критерия оценки результатов работы программного обеспечения. Рассмотрим проблему выбора этого критерия на примере задачи сегментации изображений, обычно решаемой в ходе компьютерного анализа изображений, поступающих с борта ДПЛА для выделения областей на изображении, существенных для принятия решений оператором ДПЛА.

Как известно [9], «хорошая» сегментация должна удовлетворять следующим содержательным требованиям:

- сегменты должны быть по некоторым характеристикам однородными;
- 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.

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

- иногда результаты сегментации, явно лучшие с точки зрения экспертов, имели более высокий процент ошибочно классифицированных пикселей;
- 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 сегментов изображения должны иметь разный вес;
- отсутствовала информация о том, какой класс пикселей вносил наибольшую ошибку.

Для решения последних двух проблем в [10] были предложены два критерия, являющиеся обобщением для случая нескольких классов ошибок первого и второго рода: доля неправильно отнесенных пикселей к общему числу пикселей, не принадлежащих сегменту, и доля ошибочно не отнесенных к сегменту пикселей к общему числу пикселей сегмента соответственно.

Еще одна мера оценки качества сегментации, основанная на подсчете неправильно классифицированных пикселей и использующая байесовский подход, была предложена в [11]. В этой работе для случая одного сегмента вычисляются оценки вероятности того, что случайно выбранный пиксель на отсегментированном изображении принадлежит сегменту и, соответственно, фону. Используя стандартные вероятностные формулы, вводится вероятность ошибки сегментации для всего изображения $p(err)$:

$$p(err) = p(o)p(b|o) + p(b)p(o|b),$$

где $p(o)$, $p(b)$ – априорные вероятности того, что случайным образом выбранный пиксель исходного изображения принадлежит сегменту или, соответственно, фону; $p(o|b)$ – вероятность того, что пиксель, принадлежащий фону, был ошибочно отнесен к сегменту; $p(b|o)$ – вероятность того, что пиксель, принадлежащий сегменту, был ошибочно отнесен к фону. Эти вероятности определяются как отношение соответствующих областей на изображении.

В работе [10] была предложена мера ε, основанная на следующем подходе:

$$\varepsilon = \frac{\sqrt{\sum_{i=1}^{N} d_i^2}}{A} \times 100,$$

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 сегменту.

Для оценки качества сегментации используется также критерий FOM_e [12], являющийся модификацией широко используемой метрики Прэтта. В этой же работе была предложена мера *FOC* (Figure Of Certainty), при вычислении которой используется информация об интенсивности изображения.

К этому же классу мер относятся критерии *AUMA* (Absolute Ultimate Measurement Accuracy) и *RUMA* (Relative Ultimate Measurement Accuracy) [13], оценивающие качество сегментации по тому, насколько точно (по сравнению с эталоном) можно после сегментации определить на изображении ряд характеристик, существенных для анализа изображения оператором.

Приведенные выше многочисленные критерии оценки качества работы программ, осуществляющих сегментацию изображения, наглядно иллюстрируют необходимость тщательного подхода к выбору критерия, который должен быть содержательно значимым для решения оператором ДПЛА возложенных на него задач, и, соответственно, являющегося определяющим при сравнении качества работы различных программ, решающих эту задачу.

Требования к тестовым материалам

Выбор тестового материала, то есть множества эталонов с ground truth, наряду с критериями качества также является элементом методики EDEM, причем ключевым для рассматриваемого случая оценки программных продуктов на основе эмпирической методики. Целью такой оценки программного продукта является исследование его поведения на различных категориях тестовых материалов (в рассматриваемом случае – тестовых изображений) и помощь в выборе наилучших параметров программы в различных ситуациях.

Обычно выделяют четыре основных способа оценки (тестирования) работы программы на тестовых изображениях [14]. Первый из них, исчерпывающее тестирование, представляет собой грубый подход к тестированию, основанный на переборе всех изображений из имеющейся базы данных. Этот подход зачастую является чрезмерным. Следующий способ – тестирование на экстремальных в некотором смысле изображениях (boundary value testing), оценивает работу программы на определенной пользователем репрезентативной выборке изображений из базы данных. Третий способ – случайное тестирование, основанное на произвольном выборе изображений из базы данных. При таком тестировании могут возникать более разнообразные ситуации, чем при тестировании на «экстремальных» изображениях, поскольку в последнем случае выбор таких изображений носит субъективный характер и может не учитывать всего многообразия случаев, возникающих на практике. Последний способ, тестирование в наихудших случаях (worst case testing), включает в себя рассмотрение ситуаций, когда изображение содержит редкие или необычные свойства.

Выбор конкретного способа тестирования программы зависит от внешних, содержательных требований, предъявляемых к качеству тестирования, и является первым требованием, предъявляемым к тестовым изображениям в целом. Следующее требование к изображениям – надежность получаемых результатов по качеству тестируемых программ. В зависимости от выбора способа тестирования это требование преобразуется в следующие требования к тестовому материалу. При выборе первого способа тестирования таким требованием будет полнота отображения в конечном наборе эталонов ситуаций, возникающих при решении конкретной практической задачи, включая условия применения. Это требование плохо формализуется, и поэтому доказательство полноты, как правило, невозможно. В таком случае ограничиваются экспертными оценками полноты тестового материала, при необходимости требуя его пополнения. При выборе второго способа, во-первых, нужно формально определить характеристики изображений, выбранных для построения репрезентативной выборки, и, во-вторых, формально или экспертно оценить достаточность их в базе изображений для построения репрезентативной выборки. В случае нехватки требуется пополнение изображениями с соответствующими «экстремальными» характеристиками. При выборе третьего способа тестовый материал должен быть представительным, другими словами, в нем должны быть представлены в достаточном с точки зрения статистики количестве все ситуации, существенные относительно цены ошибок. Как и ранее, полная формализация этого требования затруднительна и проверка его обычно осуществляется экспертно. При нарушении требования представительности необходимо пополнение. Требования к тестовому материалу в случае четвертого способа тестирования аналогичны требованиям, предъявляемым при выборе второго способа тестирования.

Как видим, требование надежности к тестовым изображениям трансформируется в требование полноты тестового материала по характеристикам, зависящим от выбора способа тестирования.

Результаты апробации элементов методики EDEM

В последние годы в НИИСИ РАН для сравнительного исследования алгоритмов обработки и анализа изображений была разработана и успешно развивается программная среда PICASSO (PICture Algorithms Study SOftware) [15]. Она была создана в качестве инструмента разработки адаптивных систем анализа изображений для широкого спектра прикладных задач, что позволило выбрать ее для проведения апробации элементов методики оценки качества EDEM. В качестве одной из задач дистанционного мониторинга для такой апробации была взята задача сегментации, решаемая в рамках так называемого boundary-based-подхода, широко применяемого на практике и основанного на результатах применения детектора границ. Следует отметить, что число таких детекторов постоянно растет, так что проблема выбора среди них оптимального для решения конкретной задачи является весьма актуальной. На рисунке 1 приведены исходное изображение и результаты обработки его тремя детекторами границ в программной среде PICASSO.

Результаты применения различных детекторов границ визуально различимы, но сложно сделать обоснованное утверждение о преимуществах того или иного детектора на основе только визуального анализа. Программная среда PICASSO позволяет осуществлять выбор критерия оценки результатов работы программы, определяемого смыслом решаемой задачи, и сформировать материал для сравнительного тестирования различных программ, решающих эту задачу. Именно эти два элемента методики

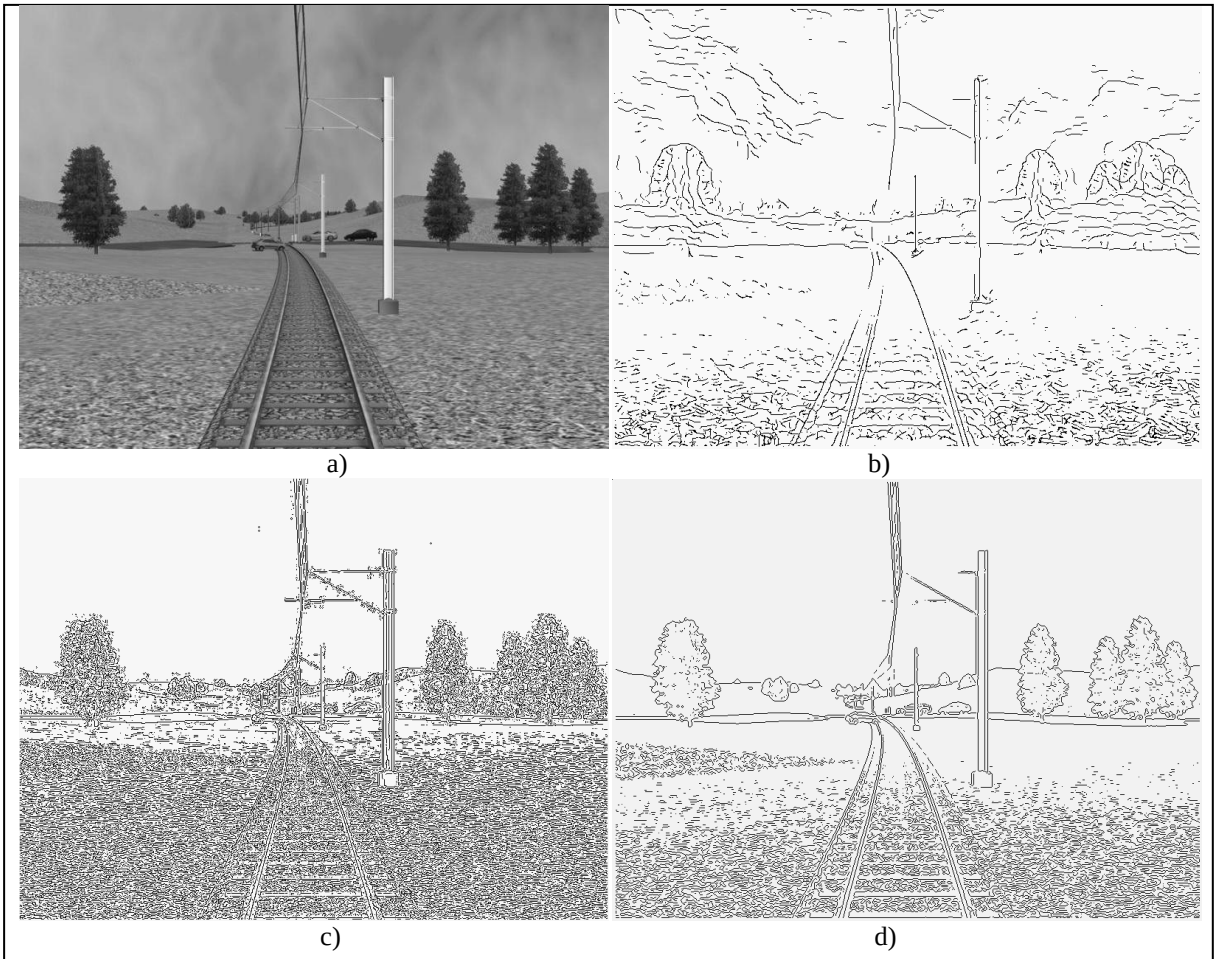

Рис. 1. Результаты обработки изображения различными детекторами границ: a) исходное изображение, b) алгоритм Heitger, c) алгоритм Black, d) алгоритм Rothwell

EDEM были апробированы для проверки возможности получения оценки качества программ. В качестве критерия оценки результатов работы детекторов границ были взяты *чувствительность* и *специфичность*, определяемые следующим образом.

Пусть идеальная граница состоит из N_t точек и некоторый детектор границ при обработке этого изображения выделил N_d точек границы, из которых N_{td} точек выделены верно. Тогда

$$\textit{чувствительность} = \frac{N_{td}}{N_t}, \quad \textit{специфичность} = \frac{N_{td}}{N_d}.$$

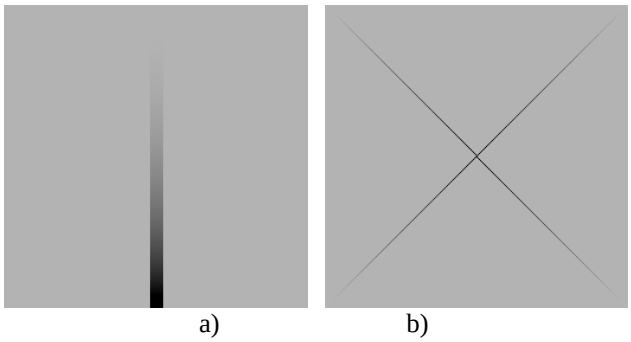

Рис. 2. Пример изображений из тестового набора системы PICASSO: a) исчезающая полоса, b) затухающий узел

В качестве тестовых материалов в программной среде PICASSO были сгенерированы искусственные изображения, моделирующие варианты наихудших для корректной работы детектора границ ситуаций. На рисунке 2 даны примеры таких тестовых изображений, сложность которых для работы детекторов границ обусловлена наличием границы изменяющегося контраста.

Набор тестовых изображений был расширен за счет добавления в него изображений, подвергнутых искажениям. В качестве искажений были взяты гауссов шум и размытие. В набор вошли изображения, полученные при разных значениях управляющих параметров

этих искажений, – величины дисперсии шума и размера окна осреднения. На рисунке 3 приведен пример полученных значений *чувствительности* и *специфичности* для детекторов Canny, Rothwell, Heitger, Black, Iverson и Smith [16], предназначенных для решения одной и той же задачи выделения границ.

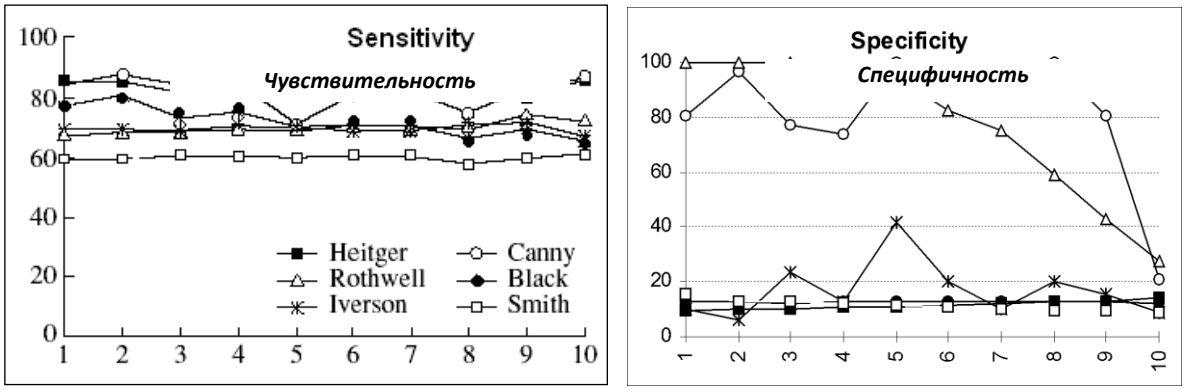

Рис. 3. Чувствительность и специфичность для тестового изображения Исчезающая Полоса. Ось ординат – чувствительность (слева) и специфичность (справа), ось абсцисс – σ-дисперсия гауссовского шума

Это исследование выявило лидеров (детекторы Canny и Rothwell) и аутсайдеров (Smith) среди тестируемых программ, подтверждая возможность использования методики EDEM для оценки качества программ. Можно отметить ее гибкость в применении к конкретным программам, позволяющую учитывать специфические особенности реализации самих программ и условий их применения. Это достигается созданием набора тестовых изображений, учитывающего все эти факторы. В качестве примера на рисунке 4 приведены изображение рабочей среды для решения задачи сегментации, одно из тестовых изображений и соответствующее ему ground truth.

Рис. 4. Рабочая среда (a) (источник – http://www.ersa-france.com), тестовое изображение (b), Ground truth (c)

В качестве второй задачи дистанционного мониторинга для апробации возможности применения методики EDEM была рассмотрена задача выделения на изображении важных в контексте решаемой внешней задачи элементов. Пример таких важных частей изображения для задачи мониторинга потенциально опасных ситуаций приведен на рисунке 5.

Принципиальная неопределенность, существующая в локализации положения границы, отделяющей объект от фона при обработке данных дистанционного зондирования, а также опыт использования программной среды PICASSO в таких условиях [16] определили использование теории нечетких множеств для оценки программных решений в задачах сегментации, а также детектирования важных элементов изображения.

Проверка возможности получения оценки свойства улавливать значимые точки на изображении выполнялось для детекторов границ Canny и Rothwell. Как и ранее, с помощью программной среды PICASSO был сделан выбор критерия оценки результатов работы программного обеспечения, а также сформирован тестовый материал. В качестве критерия оценки были взяты два варианта так называемой полной меры точности [17], обычно применяемые для определения степени принадлежности объекта

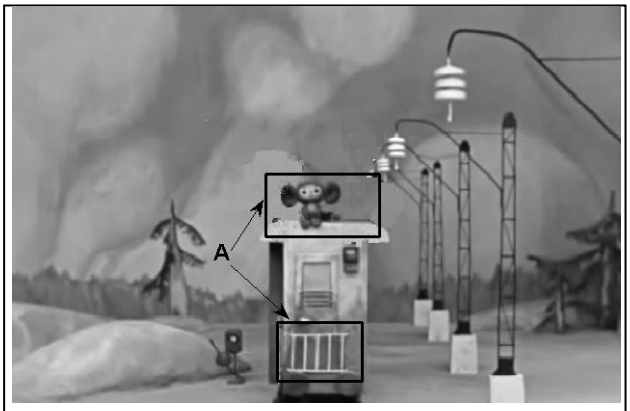

Рис. 5. Важные части изображения (A)

отрезку [0, 1] при наличии более двух нечетких множеств. В качестве тестового изображения был рассмотрен квадрат, изображенный на рисунке 6a, в котором угловые точки объявлены важными элементами. Как известно [16], детектор Canny выделяет угловые точки в отличие от детектора Rothwell. В то же время стандартные оценки качества на этом изображении для этих детекторов совпадают. Это обстоятельство преодолено введением нечеткого ground truth, изображенного на рисунке 6b. Здесь граничные пиксели квадрата разделены на два класса: вертикальные и горизонтальные. Всем граничным пикселям, принадлежащим только этим классам, функции принадлежности предписывают значение 0,8. Угловым пикселям, принадлежащим обоим классам, предписывается значение 1.

Сравнение результатов работы детекторов с использованием нечеткого ground truth продемонстрировало явное предпочтение результатов работы детектора Canny по обоим вариантам полных мер точности относительно детектора Rothwell, что вполне согласуется со здравым смыслом. Таким образом, использование методики EDEM для получения оценки качества программных продуктов возможно и для задач выделения на изображении важных элементов.

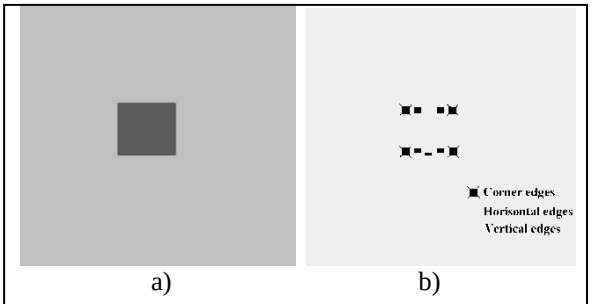

Рис. 6. Тестовое изображение (a), нечеткое ground truth (b)

В последние годы проблеме сравнительного исследования качества работы программных алгоритмов, предназначенных для решения некоторой определенной научно-технической задачи, уделяется все возрастающее внимание. Ведущая роль здесь принадлежит развитию эмпирических методик, позволяющих сделать объективный выбор на основе использования для оценки результатов работы программ тестовые наборы, содержащие известные ground truth решения задачи. При этом для оценки результатов используются различные критерии качества.

Имеющийся опыт применения объективной прямой эмпирической методики оценки качества программных продуктов, накопленный за десятилетие развития системы PICASSO, дал основание для использования лежащей в основе этой системы методики EDEM при решении проблемы объективного выбора эффективного программного обеспечения, предназначенного для решения задач дистанционного мониторинга в режиме реального времени, в том числе относящихся к области эффективного управления инфраструктурой железнодорожного транспорта. В рамках этого выбора была продемонстрирована широкая вариабельность критериев эффективности, а также сформулированы требования к тестовым материалам, используемым при сравнении результатов работы программ. С помощью системы PICASSO было получено подтверждение возможности использования методики EDEM для объективной оценки качества программных продуктов с целью выбора эффективных программных решений в области обработки результатов дистанционного зондирования, показана эффективность рассмотренного подхода, определяющая перспективность его использования.

Опыт применения методики EDEM в рамках системы PICASSO показал эффективность и возможности широкой адаптации объективной прямой эмпирической методики оценки качества программных продуктов, использованной в системе, как через вариации критериев качества, так и через внесение контролируемых искажений и изменений в тестовый материал.

Авторы выражают глубокую благодарность за помощь и полезные обсуждения в ходе работы над статьей коллегам: Лисице А.В., Рудаковой Е.И., Черепнину А.А., Чеховичу Ю.В.

Литература

1. Lodemann M., Luttenberger N. Ontology-based railway infrastructure verification – planning benefits. Proc. of the Intern. Conf. on Knowledge Management and Information Sharing (KMIS 2012), Valencia, Spain, October 2010, pp. 176–181.

2. German railway operator deploys drones in war on graffiti artists. URL: http://www.businessweek.com/articles/2013-05-29/german-railway-operator-deploys-drones-in-war-on-graffiti-artists (дата обращения: 17.02.2014).

3. UAVs may be used to stop jumbos deaths on rail tracks. URL: http://archive.indianexpress.com/news/uavs-may-be-used-to-stop-jumbos-deaths-on-rail-tracks/1214917 (дата обращения: 17.02.2014).

4. Gribkov I.V., Koltsov P.P., Kotovich N.V., Kravchenko A.A., Kutsaev A.S., Nikolaev V.K., Zakharov A.V. Multicomputer System DEDAL-2 for Local Landscape Monitoring. Journ. of WSCG, 2002, vol. 10, no. 3, pp. 169–174.

5. Российские технологические платформы: Высокоскоростной интеллектуальный железнодорожный транспорт. URL: http://www.hse.ru/org/hse/tp/transp_hispeed (дата обращения: 17.02.2014).

6. Zhang H., Fritts J.E., Goldman S.A. Image segmentation evaluation: A survey of unsupervised methods. Computer Vision and Image Understanding, 2008, vol. 110, no. 2, pp. 260–280.

7. Cardoso J.S. and Corte-Real L. Toward a Generic Evaluation of Image Segmentation. IEEE Transactions on Image Processing, 2005, vol. 14, no. 11, pp. 1773–1782.

8. Zhang Y.J. A survey on evaluation methods for image segmentation. Pattern Recognition, 1996, vol. 29, no. 8, pp. 1335–1346.

9. Haralick R.M. and Shapiro L.G. Image segmentation techniques. Computer Vision, Graphics, and Image Processing, 1985, vol. 29, no. 1, pp. 100–132.

10. Yasnoff W.A., Mui J.K. and Bacus J.W. Error measures for scene segmentation. Pattern Recognition, 1977, vol. 9, no. 4, pp. 217–231.

11. Van Droogenbroeck M. and Barnich O. Design of Statistical Measures for the Assessment of Image Segmentation Schemes. Proc. of 11th Intern. Conf. on Computer Analysis of Images and Patterns (CAIP2005), Lecture Notes in Computer Science, 2005, vol. 3691, pp. 280–287.

12. Strasters K.C. and Gerbrands J.J. Three-dimensional image segmentation using a split, merge and group approach. Pattern Recognition Letters, 1991, vol. 12, no. 5, pp. 307–325.

13. Zhang Y.J. and Gerbrands J.J. Objective and quantitative segmentation evaluation and comparison. Signal Processing, 1994, vol. 39, no. 1–2, pp. 43–54.

14. Wirth M.A. Performance evaluation of image processing algorithms in CADe, Technology in Cancer Research and Treatment, 2005, vol. 4, no. 2, pp. 159–172.

15. Грибков И.В., Захаров А.В., Кольцов П.П., Котович Н.В., Кравченко А.А., Куцаев А.С., Осипов А.С. Тестирование методов сегментации изображений в системе PICASSO. М.: Изд-во НИИСИ РАН, 2007.

16. Захаров А.В., Кольцов П.П., Котович Н.В., Кравченко А.А., Куцаев А.С., Осипов А.С. Некоторые методы сравнительного исследования детекторов границ: Тр. НИИСИ РАН. 2012. Т. 2. № 1. С. 4–14.

17. Jäger G., Benz U. Measures of Classification Accuracy Based on Fuzzy Similarity. IEEE Trans. On Geoscience and Remote Sensing, 2000, vol. 38, no. 3, pp. 1462–1467.