Некоторые фундаментальные вопросы эмпирического оценивания систем компьютерного зрения*

***П. П. Кольцов*[1], *А. С. Осипов*[2], *Р. М. Сотнезов*[3], *Ю. В. Чехович*[1], *Д. А. Якушев*[4]**

kppkpp@mail.ru; osipa68@yahoo.com; sotnezov@forecsys.ru; chehovich@forecsys.ru; D.Yakushev@gismps.ru

[1]Федеральный исследовательский центр Информатика и Управление Российской Академии Наук, 119333, Москва, Вавилова, д.44, кор.2; [2]Научно-исследовательский институт системных исследований Российской Академии Наук, 117218, Москва, Нахимовский просп., 36, к.1; [3]ЗАО Форексис, 119333, Москва, ул. Вавилова 42, оф. 152; [4]ЗАО «Транспутьстрой», 107078, г. Москва, Орликов переулок, д. 5, стр. 2.

Статья посвящена вопросам сравнительного исследования алгоритмов обработки и анализа изображений, используемых при создании различных программно-технических средств обеспечения безопасности. Изложены основные принципы разработанной для этой цели методологии EDEM, при этом особое внимание уделено используемым при сравнительной оценке алгоритмам элементов теории нечётких множеств. Рассмотрены концепции нечётких ground truth образов и нечётких мер сходства. Приведены примеры использования методологии EDEM, в том числе для оценки алгоритмов решения некоторых задач обеспечения железнодорожной безопасности.

Ключевые слова: *компьютерное зрение; сравнительное исследование; ground truth образы; нечёткие множества*

DOI: 10.21469/22233792.4.1.03

1 Введение

Различные системы компьютерного зрения играют всё большую роль в решении разнообразных прикладных задач, являясь важным инструментом повышения производственной эффективности. Так, обеспечение комплексной многоуровневой безопасности движения является одной из стратегических целей инициированной ОАО РЖД технологической платформой «Высокоскоростной интеллектуальный железнодорожный транспорт» [1]. При этом ключевую роль в обеспечении железнодорожной безопасности играют разнообразные системы, использующие компьютерное зрение: от систем мониторинга состояния железнодорожных путей, до систем идентификации персонала по видеоизображению. Разнообразие программно-алгоритмических реализаций систем компьютерного зрения, призванных решать некоторую практическую задачу ставит перед практическим пользователем непростую задачу выбора системы, наиболее подходящей для его конкретных целей. При этом общая стоимость программно-технических средств обеспечения безопасности, использующих компьютерное зрение, может быть весьма высока. Например, система распознавания лиц ForensicaGPS [2], основанная на преобразовании двумерной фотографии или видеокадра в 3D образ, была приобретена спецслужбами Саудовской Аравии и, по некоторым данным, используется в создании программного обеспечения «высокотехнологичной ограды» на границе с Ираком. При этом общая стоимость создания

*Работа выполнена при финансовой поддержке РФФИ, проект № 17-20-02205.

такого ограждения оценивается в 3.4 млрд. евро. Зачастую надёжность предлагаемых программных продуктов для решения конкретной практической задачи недостаточно высока. Всё вышеперечисленное делает проблему сравнительной оценки данных программно-алгоритмических реализаций решения некоторой фиксированной задачи с целью выявления лучших (для конкретного практического применения) весьма острой [3, 4]. Очевидно, только получив объективную оценку качества различных решений, можно определить среди них наиболее эффективное с практической точки зрения. Именно эта проблема и пути ее решения легли в основу методики сравнительной оценки программных средств в области обработки и анализа изображений EDEM [4–6], особенности которой будут рассмотрены в следующем разделе.

2 Методология EDEM: основные свойства

Прежде всего, следует отметить, что к настоящему времени не выработано единой методики оценки качества работы различных компьютерных программ, решающих некоторую содержательную задачу в области обработки и анализа изображений. Основные отличия методик, применяемых в сравнительных исследованиях различных подобных программ, заключаются в используемом критерии оценки (количественный или качественный, использующий эталонное решение или нет), использованном типе эталонных изображений (реальные или синтезированные), их количестве, параметрах, источниках (общедоступные или оригинальные) и т. п.

Известно несколько попыток классифицировать эти методики. Так, в работе [7], была использована достаточно популярная среди исследователей классификация методик сравнительного исследования алгоритмов сегментации изображений (вполне применимая и к другим классам алгоритмов компьютерного зрения), согласно которой методики оценки делятся на субъективные и объективные. Первые из них ориентированы на получение оценок качества на основе мнения экспертов. 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. Методика EDEM относится к разряду прямых методик, ориентированных на получение оценки качества работы программной реализации алгоритма, решающего конкретную задачу из области обработки и анализа изображений. Можно отметить, что развитие концепций, используемых в прямых объективных методиках (прежде всего, концепций эталонов и метрик) полезно и для системных методик оценки. Среди прямых объективных методик различают аналитические и эмпирические [7]. Аналитические методики рассматривают алгоритм независимо от его выхода, при этом исследуются такие свойства алгоритма, как стратегия реализации главной цели, сложность, ресурсоемкость и т.п. В свою очередь, эмпирические методики, оценивают не сам алгоритм, а результаты его работы на некотором наборе тестовых изображений. Такие методики с помощью вариации изображений позволяют оценить качество работы компьютерных программ на широком спектре внешних условий с учётом особенностей практического применения программ, включая границы применимости. Методика EDEM строит оценку качества на основе количественной меры различия между результатами работы программы на некотором наборе изображений, для которых точное решение, так называемое ground truth, известно априори, и этими точными решениями. 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 обстоятельства определили выбор названия нашей методики: EDEM (Empirical Discrepancy Evaluation Method).

Первоначальные опыты по сравнительному оцениванию детекторов границ и алгоритмов сегментации изображений позволили сформулировать базовые принципы используемого нами оценивания, составляющие суть методологии EDEM (подробнее см. [4]):

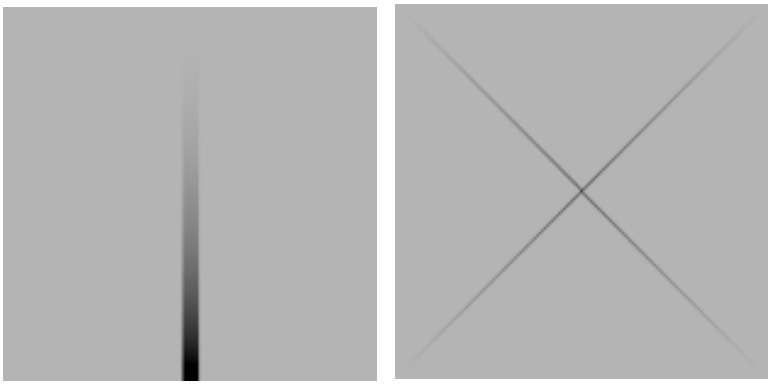
1. Оценивание использует априорное знание решения частных задач, которых должно решать оцениваемое средство.
2. Набор частных задач с известным истинным решением должен быть представительным как с точки зрения сравниваемых средств, так и с точки зрения ожидаемых условий применения.
3. Оценивание проводится на основе меры близости между результатом работы оцениваемого средства на наборе частных задач и истинным решением задачи.
4. Выбор конкретных мер близости делается априори исходя из требуемой содержательной интерпретации результатов оценивания.

Сформулированные принципы, в свою очередь, позволили построить следующую технологическую цепочку для работы по сравнительному анализу программных средств в области обработки визуальной информации [4]:

1. Определение, какие конкретно свойства алгоритмов и их программных реализаций, и при каких условиях будут тестироваться (задание глубины тестирования).
2. Задание набора эталонов – ground truth изображений и тестовых изображений, удовлетворяющих требованиям по глубине тестирования.
3. Определение меры близости между результатами работы тестируемых средств по всему набору тестовых изображений и истинными ground truth решениями.
4. Построение ранжированной на основе значений мер близости последовательности сравниваемых программных средств.
5. Интерпретация на основе значений использованных мер близости результатов ранжирования.

При выборе тестовых изображений необходимо учитывать как специфические особенности реализации самих программ, так и условия их применения. При этом изображения должны содержать ситуации, трудные для тестируемых алгоритмов. Применительно к исследованию детекторов границ, соответствующие примеры приведены рис. 1. Здесь сложность ситуации для работы детекторов границ обусловлена наличием границы изменяющегося контраста.

Первоначально, при оценке детекторов границ и алгоритмов сегментации, в качестве мер близости были использованы статистические меры оценки качества, основанные на проценте правильно классифицированных пикселей тестового изображения. Применительно к детекторам границ это может быть процент правильно (в терминах ground truth решений) определённых граничных пикселей, а применительно к методам сегментации – процент пикселей, отнесённых методом к «правильным» сегментам. Наиболее слабым местом статистических мер является то, что различия между изображениями A и B определяются по общему числу расхождений между ними, безотносительно к образу, который эти изображения представляют. 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. В результате, оценки программных средств, опирающиеся только на значения статистических показателей, вполне могут противоречить здравому смыслу



(*а*) Исчезающая полоса (*б*) Затухающий узел

Рис. 1 Пример тестовых изображений для исследования детекторов границ

(соответствующие примеры приведены в [6]). Кроме того, углубленное тестирование алгоритмов на основе значений таких мер (например, применительно к оценке детекторов границ, когда оценивается способность алгоритмов к выделению слабоконтрастных границ, или их способность к выделению непрерывных границ) оказывается затруднительным как по подбору тестовых изображений и соответствующих ground truth эталонов, так и по анализу и интерпретации результатов. Таким образом, возникает естественный подход, заключающийся в сочетания мер, относящихся к различным классам. Так, для тестирования детекторов границ и алгоритмов сегментации, были использованы сочетание статистических мер и мер оценки качества локализации (также имеющих свои недостатки), например, метрики Пратта (подробнее см. [6]). Кроме того, для более глубокого тестирования, помимо тестовых изображений, моделирующих сложные для распознавания ситуации, оказалось эффективным использование их упрощённые аналогов. В качестве примера, на рис. 2 приведено упрощённое изображение, соответствующее образу Исчезающая полоса (рис. 1, *а*), используемое при тестировании детекторов границ. Здесь контраст границы постоянен. Поскольку ряд известных алгоритмов обработки и анализа изображений имеет многочисленные программные реализации, актуальной является проблема отделения тестирования собственно алгоритмов от тестирования их программных реализаций. Для выявления ошибок последних, в рамках методологии EDEM, предусмотрен ряд простых тестов, составленных на основе специфики рассматриваемой задачи. Например, при исследовании поведения детекторов границ при аффинных преобразованиях (сдвигах, поворотах, сжатиях/растяжениях) объектов на исходных изображениях, поведение алгоритмов тестируется при повороте объекта на изображении на 180 градусов. Очевидно, результаты тестирования алгоритма на исходном и повернутом изображении должны быть близкими к идентичным (подробнее см. [6]).

Несовершенство съемочной аппаратуры, неопределённость, существующая в локализации положения границы, отделяющей объект на изображении от фона, а также другие объективные причины свидетельствуют в пользу перспективности применения теории нечётких множеств в задачах обработки и анализа изображений. В последние десятилетия для решения подобных задач был разработан ряд алгоритмов, использующих элементы нечёткой логики. Соответственно, методология EDEM также включает в себя эти элементы, что позволяет одновременно оценивать традиционные «четкие» и «нечёткие»

(*а*) Ground truth образ (*б*) Упрощенная версия

Рис. 2 Изображения, соответствующие рис. 1, *а*

алгоритмы между собой и сравнивать между собой результаты их оценки. Кроме того, «нечёткие» элементы, содержащиеся в данной методологии, делают тестирование и тех и других алгоритмов более глубоким. Именно, в рамках данной методологии используются две концепции: нечетких мер сходства (в частности, такие меры позволяют сравнивать между собой четкие и нечеткие множества) и нечетких ground truth образов. Эти концепции будут рассмотрены ниже подробнее. В работе [9], посвященной исследованию детекторов границ, было впервые предложено использовать различные нечеткие ground truth образы, соответствующие одному и тому же тестовому изображению. Это позволило лучше протестировать те или иные свойства рассматриваемого детектора границ. Оказалось, что, некоторые такие образы лучше использовать для проверки способности тестируемого детектора выделять слабоконтрастные границы, в то время как другие более приспособлены для проверки способности детектора к выделению непрерывных границ. Другим важным приложением нечетких ground truth образов, содержащимся в [9], является возможность их использования при исследовании свойства детектора границ к выделению граничных точек, существенных для определения ограничиваемого объекта (англ. image feature points). Например, для прямоугольника таковыми являются угловые точки. Применительно к оценке алгоритмов сегментации изображений, использование нечёткой логики рассмотрено в [4,5], при этом подход предложенный к оценке детекторов границ, оказался применимым и в этом случае.

Таким образом, объективная методология эмпирической оценки алгоритмов обработки и анализа изображений EDEM обладает следующими основными особенностями:

1. Использование тестовых изображений, моделирующих трудные для алгоритма ситуации, включая внесение в тестовые изображения контролируемых искажений.
2. Сочетание сложных тестовых изображений с их упрощёнными версиями.
3. Использования метрик разных классов для количественной оценки качества алгоритмов, в том числе и сочетание статистических мер с мерами, оценивающими некоторые важные свойства тестируемых алгоритмов.
4. Организация процесса сравнительного тестирования, дающая возможность качественного анализа его результатов, включая построения графиков для сравнительного их анализа.
5. Использование для тестирования элементов теории нечётких множеств, в том числе включая разработку нечётких ground truth образов с использованием нескольких

таких образов, соответствующих одному тестовому изображению, для анализа различных свойств тестируемого алгоритма.

В настоящее время, «нечёткая» компонента методологии EDEM находится в состоянии дальнейшего развития и практической апробации. В следующих разделах данной статьи будут изложены полученные в ходе исследований результаты в части решения задачи оценки программно-технических средств обеспечения безопасности, ориентированных на использование систем компьютерного зрения, в том числе и в области обеспечения безопасности на железнодорожном транспорте.

3 Нечёткие меры и ground truth образы

Существенное влияние на развитие методологии EDEM оказала работа [10], где было проведено обобщение нескольких известных статистических мер оценки качества 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 данной работы предназначались для анализа аэрокосмических снимков (определения на них областей лесов, воды, городской застройки и т. п.). Оказалось, что результаты данной работы могут быть применены для исследования производительности различных видов алгоритмов обработки и анализа изображений (в частности, детекторов границ). Это нашло своё отражение в работе [9], где несколько известных («чётких») детекторов границ было протестировано с использованием нечётких мер сходства и нечётких ground truth образов. Прежде, чем описать результаты применения «нечёткой» компоненты методологии EDEM, следует напомнить несколько основных понятий из теории нечётких множеств.

Именно, пусть X есть непустое множество (например, множество пикселей изображения). Нечёткое множество C на X есть пара $< X, f_C >$, где f_C есть отображение X на $[0, 1]$. Значение $f_C(x)$ для элемента $x \in X$ называется степенью принадлежности x множеству C (например, степень принадлежности данного пикселя тестового образа множеству границ эталонного ground truth образа), а функция f_C называется функцией принадлежности нечёткого множества. Нечёткое множество называется непустым, если хотя бы для одного элемента $x \in X$, $f_C(x) > 0$.

Заметим, что обычные (чёткие) подмножества M из X включаются в данный подход, если мы будем рассматривать их как стандартные характеристические функции $1_M : X \to [0, 1]$. То есть, например, если у нас имеется пиксель x, относящийся к некоторому классу C, то в этом случае $f_C(x) = 1$ и $f_{C_1}(x) = 0$ для всех классов C_1 отличных от C. В дальнейшем будем предполагать, что множество X конечно.

Нечёткой классификацией F множества X

$$F :=< X, f_{C_1}, \ldots, f_{C_N} >, \quad f_{C_m} : X \to [0, 1], \quad m = 1, \ldots, N$$

называется совокупность N нечётких множеств (классов) $< X, f_{C_m} >$, удовлетворяющих для любого $x \in X$ условию:

$$\sum_{m=1}^{N} f_{C_m}(x) = 1.$$

Функция f_{C_m} — степень принадлежности соответствующему классу. Легко видеть, что обычная классификация X (т. е. разбиение X на N непересекающихся подмножеств-классов) является нечёткой классификацией (каждый элемент X принадлежит ровно од-

ному классу), и в роли функций данной классификации выступают характеристические функции классов.

Обозначим множество всех нечётких множеств на X за $[0,1]^X$. Для A и B из $[0,1]^X$ нечёткое отношение включения

$$A \subset B \text{ означает, что } f_A(x) \leqslant f_B(x) \text{ для всех } x \in X.$$

Нечёткая мера сходства есть отображение $s : [0,1]^X \times [0,1]^X \to [0,1]$, сопоставляющее множествам A, $B \in [0,1]^X$ степень сходства $s(A,B) \in [0,1]$, удовлетворяющее условиям:

– $s(A,A) = 1$ для любого нечёткого A,
– $s(A,C) \leqslant s(A,B) \wedge s(B,C)$ при $A \subset B \subset C$,

где $p \wedge q$ обозначает минимум из p и q; максимум из p и q обозначается как $p \vee q$. Последнее условие представляет собой аналог неравенства треугольника для обычных метрик. Мера называется симметрической, если дополнительно

– $s(A,B) = s(B,A)$ для всех нечётких A и B.

Важными для дальнейших рассмотрений примерами данных мер являются:

$$S_1(A,B) = \frac{\sum_x (f_A(x) \wedge f_B(x))}{\sum_x (f_A(x) \vee f_B(x))},$$

$$S_2(A,B) = \frac{2\sum_x (f_A(x) \wedge f_B(x))}{\sum_x (f_A(x) + f_B(x))}.$$

Легко видеть, что они удовлетворяют всем трём приведённым выше условиям. Как показано в [10], для обычных чётких множеств A и B (отождествлённых со своими характеристическими функциями), S_1 и S_2 совпадают с мерами оценки качества классификации Шорта и Хеллдена соответственно (Short's and Hellden's classification accuracy measures). Изначально эти меры определялись только для чётких множеств, так что можно рассматривать меры S_1 и S_2 как их обобщение.

На базе мер S_1 и S_2, в работе [10] также введены итоговые (по всем классам) меры сходства (overall accuracy measures):

$$OA_1(F^1,F^2) = \frac{\sum_m \sum_{x\in X} f^1_{C_m}(x) \wedge f^2_{C_m}(x)}{\sum_m \sum_{x\in X} f^1_{C_m}(x) \vee f^2_{C_m}(x)},$$

$$OA_2(F^1,F^2) = \frac{\sum_m \sum_{x\in X} 2(f^1_{C_m}(x) \wedge f^2_{C_m}(x))}{\sum_m \sum_{x\in X} f^1_{C_m}(x) + f^2_{C_m}(x)}, \quad m = 1,\ldots,N.$$

Что касается нечётких ground truth образов, к настоящему времени не выработано общих правил их построения. В [10] было сделано предположение, что их включение в системы сравнительного тестирования детекторов границ позволит не только тестировать нечёткие детекторы, но сделать процедуру тестирования всех детекторов более содержательной. В [9] было предложено использовать различные нечёткие ground truth образы, соответствующие одному и тому же тестовому изображению. Их отличают различные функции принадлежности пикселей множеству границ. Предполагалось, что они позволят лучше протестировать те или иные свойства исследуемого детектора границ. При этом, значения функций принадлежности могут зависеть от различных факторов: расположения граничных пикселей, контраста границы и т. п. Эксперименты в частности, показали,

что одни нечёткие ground truth образы лучше использовать для проверки свойства тестируемого детектора выделять непрерывные границы, в то время как другие - для проверки способности выделять слабо контрастные границы (см. [9, 11]).

В работах [12, 13] была предложена процедура построения нечётких ground truth образов для оценки алгоритмов распознавания лиц. Там же был введён ряд новых нечётких мер сходства, обобщающих известные меры оценки качества классификации на нечёткий случай, включая упоминавшуюся выше метрику Пратта. Следующий раздел посвящённом, главным образом, результатам применения «нечёткой» составляющей методологии EDEM.

4 Практическое применение

Распознавание лиц — одна из типичных задач классификации и распознавания образов. Под задачей распознавания лиц обычно понимается следующее: дана база изображений (фотографий или видеокадров) конкретных людей, состоящая из конечного набора классов. Каждый класс изображений представляет собой изображения одного человека. По предъявлении входного изображения человеческого лица требуется определить его принадлежность одному из данных классов (или установить отсутствие такой принадлежности).

Актуальность задачи распознавания лиц, прежде всего для создания комплексных систем обеспечения безопасности, привела к появлению разных алгоритмов ее решения. Данное обстоятельство сделало актуальной проблему сравнения эти алгоритмы между собой.

Обычно образы базы изображений представляют собой обучающую выборку, при помощи которой определяются исходные параметры алгоритма распознавания. Затем, в качестве первичной верификации, к распознаванию предъявляется набор изображений тех же людей, чьи изображения содержатся в базе. Этот набор изображений составляет тестовую выборку.

Несмотря на обилие алгоритмов распознавания лиц, обычно эти алгоритмы реализованы по следующей схеме. На первом этапе решения данной задачи, для каждого изображения обучающей выборки составляется набор (вектор) его характерных признаков. На следующем этапе, используя векторы характерных признаков, с помощью алгоритмов машинного обучения строится модель — классификатор, эффективно разделяющая между собой наборы признаков, соответствующие изображениям разных лиц. Далее, для тестового изображения составляется вектор характерных признаков, который подаётся на вход классификатора, сравнивается с его содержимым, после чего делается заключение о возможной принадлежности входного вектора одному из классов обучающей выборки.

В рамках данного подхода, естественным методом классификации изображений является метод поиска ближайшего соседа в пространстве образов. Каждый образ тестовой выборки представляет собой элемент в пространстве образов и ищется ближайшее расстояние между ним и элементами обучающей выборки. Элемент тестовой выборки относят к тому же классу, что и элемент обучающей выборки, на котором достигается это расстояние. Для эффективного распознавания важно, чтобы образы, соответствующие одному классу, располагались плотно, в кластерах, и при этом кластеры, соответствующие разным классам, были разделены между собой.

Первоначально, в рамках методологии EDEM, тестирование алгоритмов распознавания лиц без использования нечёткой логики было проведено в [3, 4]. При создании тестового набора изображений в качестве эталона/ground truth образа было взято изображение лица в анфас. Затем, с помощью свободной версии программы FaceGen Modeller 3.5, по

данному изображению создавалась трехмерная модель головы, по которой создавались тестовые изображения, состоящие из изображений в профиль, а также изображений с поворотом головы и внесением контролируемых искажений (зашумления, размытия, изменением фона и т. п.). Для оценки качества работы алгоритмов было взято число правильно опознанных изображений (т. е. использовалась стандартная статистическая мера оценки качества). Таким образом, в данных экспериментах использовались такие элементы методологии EDEM как сравнение тестовых изображений с эталоном и внесение контролируемых искажений в тестовые изображения.

Как в случае детекторов границ или алгоритмов сегментации, где результатом работы алгоритма является отнесение каждого пикселя изображения к одному (в случае «чёткого» алгоритма) или нескольким (в «нечётком» случае) классам (к классу границ или фона в случае детекторов границ, или к различным сегментам в случае алгоритмов сегментации), так и в случае алгоритмов распознавания лиц результатом является классификация (как правило однозначная) каждого изображения из тестового набора. Поэтому методика оценки алгоритмов с использованием нечёткой логики, рассмотренная в предыдущем разделе, применима и для алгоритмов распознавания лиц.

Это наблюдение нашло своё отражение в [12]. В данной работе, для экспериментов была выбрана известная база изображений ORL. Она включает 400 образов, содержащих изображения лиц 40 людей (по 10 изображений каждого человека) [14]. Все изображения сделаны в полутоновом режиме, при незначительной вариации освещённости и отличаются выражением лица, поворотами головы и другими деталями.

Также для экспериментов нами был создан ряд новых изображений на основе образов базы ORL. Для этого использовалась свободно распространяемая программа морфинга изображений Sqirlz Morph версии 2.1. Она позволяет по двум изображениям лица осуществлять преобразование начального изображения лица в конечное изображение с сохранением промежуточных результатов. Пример работы данной программы приведён на рис. 3, где она использовалась для получения изображений поворотов на фиксированный угол исходного изображения лица.

Рис. 3 Исходное изображение базы ORL, его повороты вправо на 15°, 10° и влево на 25°

В качестве тестируемых алгоритмов распознавания лиц в основном тестировались демонстрационные версии программ, доступные на сайте `http://www.advancedsourcecode.com/` посвящённом программному обеспечению в области компьютерного зрения и биометрии. Все алгоритмы реализованы в среде MATLAB.

Перед началом распознавания были определены нечёткие ground truth образы. Именно, каждому элементу x тестовой выборки X (куда могли входить и элементы обуча-

ющей выборки алгоритма) предписывались априори определенные степени принадлежности каждому из распознаваемых классов. Именно, для каждого $x \in X$ определялись $g^1_{C_1}(x), \ldots, g^1_{C_N}(x)$ так, чтобы $G =< X, g_1, \ldots, g_N >$ являлась нечёткой классификацией X. Все тестируемые алгоритмы выдавали однозначный результат распознавания. Соответственно, для выборки X, элементы её тестовой нечёткой классификации $F = =< X, f_1, \ldots, f_N >$ определялись следующим образом: $f^2_{C_i}(x) = 1$ если по результатам распознавания x отнесли к классу C_i и $f^2_{C_j}(x) = 0$ для j отличных от i.

По завершении тестирования вычислялись меры OA_1 и OA_2, определённые выше, и кроме того, как и в [3, 4], доля правильно распознанных изображений (стандартная статистическая мера). При этом был получен ряд содержательных результатов. Например, тестировалась устойчивость распознавания при небольших поворотах головы распознаваемого объекта (как на рис. 3). При сравнении между собой алгоритмов Eigenface и lbp общий процент правильно распознанных изображений был выше у второго алгоритма. При использовании нашей методики, изображениям поворотов головы априорно давалась более высокая степень принадлежности (равная 1), в сравнении с остальными изображениями (степень принадлежности равна 0.8). Соответственно, правильно распознанные повороты головы оцениваются выше, чем остальные правильно распознанные изображения. В результате, значения OA_1 и OA_2 для алгоритма Eigenface оказались выше, чем для lbp-алгоритма. Таким образом, несмотря на более низкую производительность первого из этих алгоритмов в сравнении со вторым, он оказался более устойчивым к распознаванию поворотов головы. По аналогичной схеме тестировалась способность алгоритмов распознавать изображения лица в очках и без очков (некоторые классы изображений ORL, в том числе класс, соответствующий рис. 3, содержат изображения обоих таких видов).

Как отмечалось выше, для эффективной работы алгоритма распознавания, большое значение имеет расположение элементов обучающей выборки в признаковом пространстве, в частности, чтобы представители разных классов были разделены между собой. Для проверки этого свойства, при помощи Sqirlz Morph из пар изображений, представляющих разные классы, были созданы наборы гибридных изображений. Здесь для каждого гибридного изображения естественно задавать его степени принадлежности обоим родительским классам: чем ближе изображение к одному из этих классов, тем степень принадлежности к этому классу выше, и наоборот, тем ниже степень принадлежности ко второму родительскому классу. Пример родительских и гибридных изображений содержится на рис. 4.

В процессе тестирования вычислялись значения мер OA_1 и OA_2. В качества примера, количественные результаты тестирования нескольких алгоритмов на наборе гибридных изображений, составленных из родительских изображений рис. 4, представлены в таблице 1.

Таблица 1 Результаты распознавания тестовых изображений, составленных из исходных изображений рис. 4

Мера	Алгоритм		
	Fisherface	Lbp	Eigenface
OA_1	0.5686	0.3582	0.5499
OA_2	0.725	0.5275	0.7151

Рис. 4 Верхний ряд — исходные изображения из базы ORL. Нижний ряд — гибридные изображения. Степени принадлежности родительским классам для гибридных изображений, соответственно — $(0.75, 0.25)$, $(0.5, 0.5)$, $(0.25, 0.75)$

При непосредственной проверке оказалось, что алгоритм Fisherface отнёс каждое изображение из набора к тому родительскому классу, степень принадлежности к которому у изображения выше. Таким образом, разделение гибридного набора было выполнено корректно. Иная ситуация оказалась у lbp – алгоритма: из 72 изображений лишь 48 были отнесены к родительскому классу с преобладающей степенью принадлежности, 4 изображения – к родительскому классу с меньшей степенью принадлежности, а остальные 20 – к другим классам. У алгоритма Eigenface разделение гибридного набора было выполнено корректно, за исключением двух изображений (со степенями принадлежности $(0.45, 0.55)$ и $(0.44, 0.56)$), отнесённых к другим классам. Таким образом, здесь значения указанных мер оказались адекватными реальной ситуации.

Вместе с тем, в [12] отмечалась желательность разработки новых нечётких мер сходства для более качественного тестирования. Так, одним из недостатков мер OA_1 и OA_2 являются относительно невысокие (т. е. далёкие от максимума, равного 1) их значения при успешной работе алгоритма (корректном или почти корректном разделении гибридного набора). Причина здесь в том, что нечёткие ground truth классы (со степенью принадлежности меньше 1) сравниваются с результатами распознавания, которые в нечёткой терминологии либо равны 1 при верной классификации, либо равны 0 в противном случае. Для преодоления этого недостатка, в [13] вместо OA_1 и OA_2 было предложено использовать меры, основанные на введённом Л. Заде операторе интенсификации контраста (contrast intensification operator) INT:

$$INT(f(x)) = \begin{cases} 2f^2(x), & 0 \leqslant f(x) \leqslant 0.5, \\ 1 - 2(1 - f(x))^2, & 0 \leqslant f(x) \leqslant 1. \end{cases}$$

«Интенсификация контраста» состоит в том, что этот оператор увеличивает числовые значения большие 0.5 и уменьшает значения меньшие 0.5. Тем самым, функции принадлежности нечётких классов делаются ближе к характеристическим функциям обычных «чётких» множеств. Именно, вместо OA_1 и OA_2 были введены однопараметрические системы мер:

$$OA_1INT^p(F^1,F^2)=\frac{\sum_m\sum_{x\in X}INT^p(f^1_{C_m}(x)\wedge f^2_{C_m}(x))}{\sum_m\sum_{x\in X}INT^p(f^1_{C_m}(x)\vee f^2_{C_m}(x))},$$
$$OA_2INT^p(F^1,F^2)=\frac{\sum_m\sum_{x\in X}2INT^p(f^1_{C_m}(x)\wedge f^2_{C_m}(x))}{\sum_m\sum_{x\in X}INT^p(f^1_{C_m}(x)+f^2_{C_m}(x))},$$

где целочисленный параметр p — степень оператора INT $(p=0,1,2\dots)$. При $p=0$ $OA_1INT^0=OA_1$, $OA_2INT^0=OA_2$. В терминах определённых таким образом мер при $p=1$ результаты, соответствующие приведённому выше примеру, представлены в таблице 2.

Таблица 2 Результаты распознавания тестовых изображений, составленных из исходных изображений рис. 4, в терминах определённых выше мер

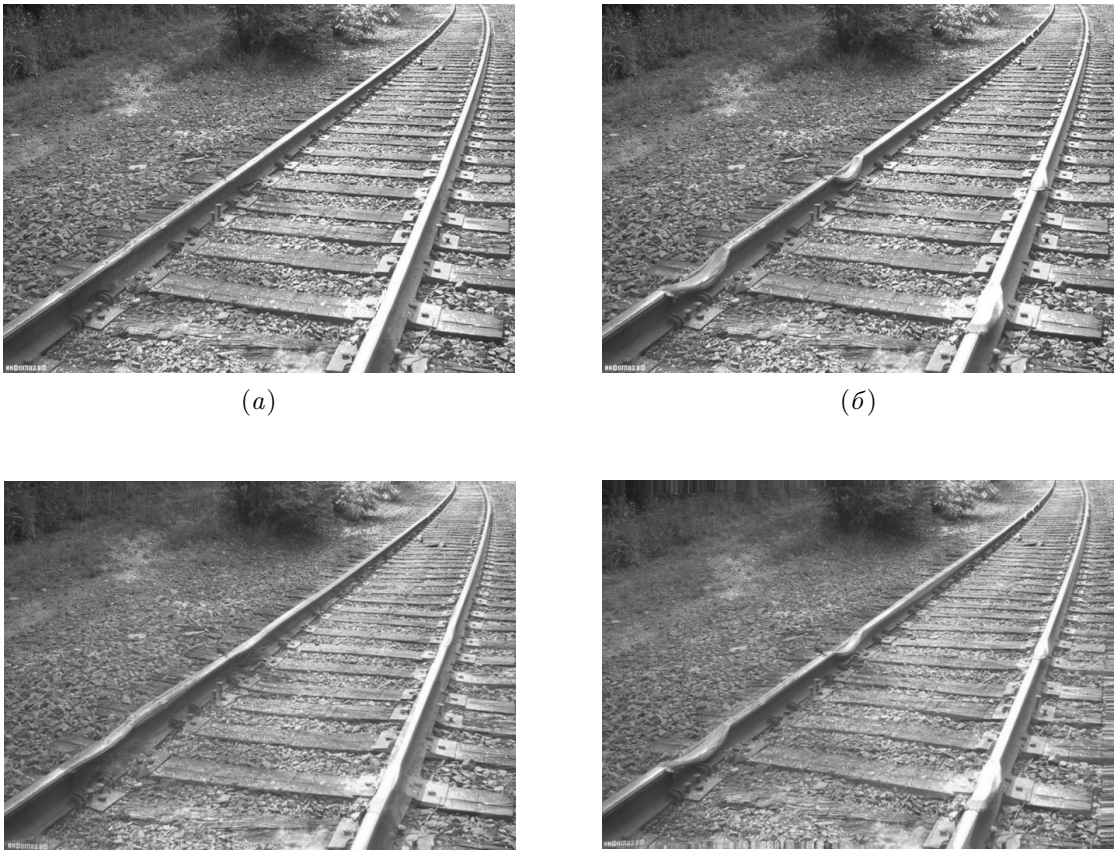
Мера	Алгоритм		
	Fisherface	Lbp	Eigenface
OA_1INT^1	0.7053	0.4213	0.6812
OA_2INT^1	0.8272	0.5929	0.8037

Сравнивая эти результаты с результатами таблицы 1, можно заметить, что увеличились как числовые значения мер, так и разности между элементами каждой строки. Последнее означает, что меры OA_1INT^1 и OA_2INT^1 более чувствительны к качеству классификации, чем меры OA_1 и OA_2.

В то же время, при тестировании свойства разделённости обучающей выборки с использованием мер OA_1INT^p и OA_2INT^p, при разных значениях параметра p ситуация иная. Так, для одного набора гибридных изображений, значения этих мер при $p=1$ для алгоритма Eigenface были выше, чем для алгоритма распознавания HOG, использующего гистограмму ориентированных градиентов изображения (подробнее см. [12, 13]). Однако при $p=2$ ситуация поменялась на противоположную. Непосредственная проверка показала, что в случае алгоритма HOG, 4 изображения гибридного набора со степенями принадлежности $(0.45,0.55),\dots,(0.42,0.58)$ были отнесены к родительскому классу с меньшей степенью принадлежности. В остальном распознавание было корректным (фактически, произошёл небольшой сдвиг в сторону одного из родительских классов). Что касается алгоритма Eigenface, то 2 гибридных изображения были отнесены алгоритмом к посторонним классам, а остальные изображения были классифицированы корректно. Таким образом, при различных значениях параметра p, эти ошибки разного типа имели разную цену. Можно сделать вывод, что исследование поведения OA_1INT^p и OA_2INT^p в зависимости от значений p остается важным открытым вопросом.

В последнее время для решения задач обеспечения железнодорожной безопасности всё активнее применяются дистанционно-пилотируемые летательные аппараты (ДПЛА) [15]. Можно отметить такие задачи с их использованием, как борьба с рисовальщиками граффити, мониторинг движения животных возле железнодорожных путей, распознавание взрывоопасных предметов на железнодорожном полотне, выявление дефектов рельсов и

опор контактной сети в полосе отвода, подробнее см. [4, 5, 15]. В работах [4, 5], на основе опыта применения методологии EDEM к оценке методов сегментации изображений, было сделано заключение о возможности применения данной методологии к задаче ДПЛА-мониторинга потенциально опасных ситуаций.



(*а*) (*б*)

(*в*) (*г*)

Рис. 5 Изображение процесса деформации рельса в результате боксования: (*а*) Исходное изображение (*б*) Деформированное изображение (*в*)-(*г*) Промежуточные изображения

Одной из таких задач является задача выявления деформации рельсов методами дистанционного зондирования. К настоящему времени предложен ряд различных алгоритмов решения этой задачи, например, измерение кривизны рельсов с использованием преобразования Хафа (Hough transform), [16]. Тем самым, и здесь возникает задача сравнительной оценки алгоритмов с целью выявления лучших. Для решения этой задачи в рамках методологии EDEM предлагается также использовать нечёткие ground truth образы и нечёткие меры сходства. Именно, вначале экспертами выбирается ряд изображений деформированных рельсов (ground truth эталонов со степенью принадлежности множества «кривых» рельсов равных 1). В качестве примера, на рис. 5, *а* изображена деформация рельсов, возникшая в результате боксования локомотива. Также, берутся соответствующие изображения недеформированных рельсов (их степень принадлежности множеству «кривых» рельсов полагается равной 0). Затем, как в случае изображений лиц, с помощью программы морфинга изображений, производится преобразование недеформированных изображений рельсов в деформированные, с сохранением промежуточных изображений

(см. рис. 5). Каждому из промежуточных изображений присваиваются степени принадлежности родительским классам (недеформированных и деформированных рельсов). Эти промежуточные изображения со степенями принадлежности своим родительским классам образуют нечёткие ground truth образы. Эти же изображения подаются на вход алгоритма, распознающего деформированные рельсы, и производится количественная оценка результатов его работы с использованием нечётких мер сходства.

Степени принадлежности промежуточных изображений родительским классам может определяться как естественным путём (аналогично примеру на рис. 4), так и на основании мнения экспертов. Методика выбора степеней принадлежности находится в стадии разработки, при этом ясно, что этот выбор должен определяться в зависимости от специфики решаемой задачи (например, насколько допустима величина деформации с точки зрения безопасности движения). В настоящее время проводится сравнительное исследование нескольких доступных методов определения кривизны рельсов с использованием рассмотренных выше нечётких мер сходства. Результаты этого тестирования составят содержание отдельной работы.

5 Заключение

При сравнительном исследовании качества работы алгоритмов компьютерного зрения, предназначенных для решения практических задач, эмпирические методики выглядят весьма перспективным, поскольку позволяют сделать объективный выбор на основе использования для оценки результатов работы исследуемых реализаций алгоритмов наборы тестовых и соответствующих им эталонных изображений, содержащие известные ground truth решения задачи. При этом использование различных количественные меры качества работы, используемое в методологии EDEM, позволяет максимально объективизировать оценку эффективности алгоритмов.

Необходимость сочетания различных мер в рамках одной методологии тестирования заключается в том, что мера производительности даёт количественную оценку, призванную характеризовать то или иное свойство тестируемого алгоритма, при этом ни одна мера не является универсальной. Таким образом, сочетание таких мер в рамках одной методологии тестирования, и разработка новых мер производительности являются важными задачами сравнительного исследования алгоритмов компьютерного зрения.

В рамках методологии EDEM разработка и анализ нечётких мер сходства (наряду с разработкой нечётких ground truth эталонов) играет важную роль. Проведённые исследования показывают, что нечёткие меры сходства, применённые в сочетании со стандартными статистическими оценками производительности, позволяют получить дополнительную информацию о тестируемых алгоритмах. Так, выше был приведён пример сравнительного исследования двух алгоритмов распознавания лиц, включающий в себя сравнение их устойчивости к поворотам головы распознаваемого объекта. В том примере статистические оценки производительности, дававшие преимущество одному алгоритму, дополнялись значениями мер OA_1 и OA_2, дававшими преимущество другому алгоритму. Это позволило сделать вывод, что хотя общая производительность первого алгоритма выше, второй из них более устойчив к малым поворотам головы объекта. Тем самым, именно различия между статистическими и нечёткими мерами сыграли здесь взаимно дополняющую роль. В то же время, использование мер OA_1INT^1 и OA_2INT^1 вместо OA_1 и OA_2 здесь нежелательно, так как легко видеть, что в этой ситуации первые из них ближе по поведению к статистическим мерам, чем вторые, и при их применении полезная дополнительная информация могла быть упущена. Таким образом, возникает важная задача

определения ситуаций, в которых использование тех или иных нечётких мер сходства оказывается оправданным. Для параметрических мер важным здесь является исследование поведения мер в зависимости от значений их параметров. С разработкой и исследованием нечётких мер сходства связана задача разработки нечётких ground truth эталонов, в частности, при оценке алгоритмов классификации, разработка методики определения степеней их принадлежности соответствующим классам. Решение этих задач составляет основное направление текущих исследований.

В целом, опыт применения методики EDEM показал эффективность и возможности адаптации объективной прямой эмпирической методики оценки алгоритмов компьютерного зрения, как через вариации количественных критериев оценки производительности, так и через внесение контролируемых искажений и изменений в тестовый материал.

Литература

[1] Технологическая платформа «Высокоскоростной интеллектуальный железнодорожный транспорт». URL: `http://www.rzd-expo.ru/innovation/technology_platform_quot_high_intellectual_rail_transport_quot/Tehnologicheskaya_platforma_2012_2015.pdf`.

[2] Facial Recognition Evolves to 3D, ForensicaGPS Unveiled. URL: `http://www.defenseworld.net/news/6981/Facial_Recognition_Evolves_to_3D__ForensicaGPS_Unveiled`.

[3] *Захаров А. В., Кольцов П. П., Котович Н. В., Кравченко А. А., Куцаев А. С., Лисица А. В., Осипов А. С., Рудакова Е. И., Черепнин А. А., Чехович Ю. В.* О методике тестирования некоторых программных продуктов в области железнодорожной безопасности // Актуальные проблемы управления перевозочным процессом. Вып. 13, ПГУПС, СПБ. 2015. С. 187–192.

[4] *Захаров А. В., Кольцов П. П., Котович Н. В., Куцаев А. С., Кравченко А. А., Осипов А. С.* О технологии сравнительного анализа программно-технических решений в области обработки визуальной информации в интересах ОАО РЖД // Сборник трудов конференции ИСУЖТ-2014. Т. 2. ОАО «НИИАС», 2015. С. 187–192.

[5] *Захаров А. В., Кольцов П. П., Котович Н. В., Кравченко А. А., Куцаев А. С., Осипов А. С.* Прямая оценка качества программных продуктов. Критерии и тестовые материалы // Программные продукты, системы и алгоритмы. № 3, 2014. С. 1–8.

[6] *Кольцов П. П., Осипов А. С., Куцаев А. С., Кравченко А. А., Котович Н. В., Захаров А. В.* О количественной оценке эффективности алгоритмов анализа изображений // Компьютерная Оптика, 2015. Т. 39. № 4. С. 542–556.

[7] *Zhang H., Fritts J. E., Goldman S. A.* Image segmentation evaluation: A survey of unsupervised methods // Computer Vision and Image Understanding, 2008. Vol. 110. № 2. P. 260–280.

[8] *Zhang Y. J.* A survey on evaluation methods for image segmentation // Pattern Recognition, 1996. Vol. 29. № 8. P. 1335–1346.

[9] *Osipov A.* A fuzzy approach to performance evaluation of edge detectors // in Lecture Notes in Signal Science, Internet and Education, WSEAS Press, 2007. P. 94–99.

[10] *Jäger G., Benz U.* Measures of Classification Accuracy Based on Fuzzy Similarity // IEEE Trans. On Geoscience and Remote Sensing, 2000. Vol. 38. № 3. P. 1462–1467.

[11] *Грибков И. В., Захаров А. В., Кольцов П. П., Котович Н. В., Кравченко А. А., Куцаев А. С., Осипов А. С.* Некоторые вопросы количественной оценки производительности детекторов границ // Программные продукты и системы, 2011. № 4. С. 13–20.

[12] *Осипов А. С.* Об использовании элементов нечёткой логики в оценке алгоритмов идентификации лиц // Труды НИИСИ РАН, 2016. Т. 6. № 2. С. 62–69.

[13] *Осипов А. С.* Нечёткие меры и их использование в оценке алгоритмов компьютерного зрения // Труды НИИСИ РАН, 2017. Т. 7. № 1. С. 46–57.

[14] База ORL изображений лиц лаборатории AT&T. URL: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[15] *Лёвин Б. А., Бугаев А. С., Ивашов С. И., Разевиг В. В.* Дистанционно-пилотируемые летательные аппараты и безопасность пути // Мир Транспорта, 2013. Т. 11. №2. С. 152–157.

[16] *Johnson C.* Image processing techniques for the detection and characterisation of features and defects in railway tracks // PhD Thesis, Manchester Metropolitan University, UK, 2013.

Поступила в редакцию 28.05.2018

Some fundamental issues of empirical evaluation for computer vision systems*

***P. P. Koltsov*[1], *A. S. Osipov*[2], *R. M. Sotnezov*[3], *Yu. V. Chehovich*[1], *D. A. Yakushev*[4]**

kppkpp@mail.ru; osipa68@yahoo.com; sotnezov@forecsys.ru; chehovich@forecsys.ru; D.Yakushev@gismps.ru

[1]Federal research center Informatics and Management of the Russian Academy of Science, 119333 Moscow Ulitsa Vavilova 44; [2]Scientific Research Institute of System Development, Nakhimovskiy Prospekt, 36/1, Moscow, 117218; [3]Address: Ulitsa Vavilova, 42, Moscow, 119333; [4]Orlikov Pereulok, 5 Moscow,107139

The paper deals with the comparative study of image processing analysis algorithms implemented in the software and hardware based security systems. The main principles of EDEM methodology, implemented for this purpose, are considered with the focus on elements of the fuzzy set theory used for the comparative evaluation. In particular, the concepts of fuzzy ground truth images and fuzzy similarity measures are considered. Some examples of application of EDEM methodology, including the evaluation of algorithms used for solving some rail security tasks are given.

Keywords: *computer vision; comparative study; ground truth images; fuzzy sets*

DOI: 10.21469/22233792.4.1.03

References

[1] Rossiiskie tehnologicheskie platformy: Visokorostnoy intellektualniy zheleznodorozhniy transport [Russian technological platforms: High speed rail transport]. URL: http://www.rzd-expo.ru/innovation/technology_platform_quot_high_intellectual_rail_transport_quot/Tehnologicheskaya_platforma_2012_2015.pdf (accessed October 30, 2017).

[2] Facial Recognition Evolves to 3D, ForensicaGPS Unveiled. URL: http://www.defenseworld.net/news/6981/Facial_Recognition_Evolves_to_3D__ForensicaGPS_Unveiled (accessed February 17, 2014).

[3] Zakharov A. V., P. P. Koltsov, N. V. Kotovich, A. A. Kravchenko, A. S. Kutsaev, A. V. Lisitsa, A. S. Osipov, E. I. Rudakova, A. A. Cherepnin, and Yu. V. Chechovich. 2015. O metodike testirovaniya nekotorih programmnih produktov v oblasti zheleznodorozhnoi bezopasnosti [On a method of testing for some rail security software]. *Actual problems of railway transport process controlling*. Petersburg State Transport University, 13:187–192.

*The research was supported by the Russian Foundation for Basic Research (grant 17-20-02205).

[4] Zakharov A. V., P. P. Koltsov, N. V. Kotovich, A. A. Kravchenko, A. S. Kutsaev, and A. S. Osipov. 2015. O tekhnologii sravnitelnogo analiza programmno-tehnicheskih reshenij v oblasti obrabotki vizual'noi informacii v interesah OAO RZhD [A technology for comparative analysis of visual processing software for the benefit of RZD Capital PLC]. *Proceedings of ISUZhT-2014 conference*. JSC NIIAS 2:187–192.

[5] Zakharov A. V., P. P. Koltsov, N. V. Kotovich, A. A. Kravchenko, A. S. Kutsaev, and A. S. Osipov. 2014. Priamaia ocenka kachestva programmnih produktov. Kriterii I testovie materialy [Direct software quality evaluation. Criteria and testing materials]. *Programmnie produkti, sistemi i algoritmi*. 3:1–8.

[6] Zakharov A. V., P. P. Koltsov, N. V. Kotovich, A. A. Kravchenko, A. S. Kutsaev, and A. S. Osipov. 2015. O kolichestvennoi ocenke effektivnosti algoritmov analiza izobrazhenij [A quantitative performance evaluation of image analysis algorithms]. *Computer Optics*. 39(4):542–556.

[7] Zhang H., J. E. Fritts, and S. A. Goldman. 2008. Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*. 110(2):260–280.

[8] Zhang Y. J. 1996. A survey on evaluation methods for image segmentation // *Pattern Recognition*. 29(8):1335–1346.

[9] Osipov A. 2007. A fuzzy approach to performance evaluation of edge detectors. In Lecture Notes in Signal Science, Internet and Education, WSEAS Press. P. 94–99.

[10] Jäger G., and U. Benz. 2000. Measures of Classification Accuracy Based on Fuzzy Similarity. *IEEE Trans. On Geoscience and Remote Sensing*. 38(3):1462–1467.

[11] Gribkov I. V., A. V. Zakharov, P. P. Koltsov, N. V. Kotovich, A. A. Kravchenko, A. S. Kutsaev, and A. S. Osipov. 2011. Nekotorye kolichestvennye ocenki proizvoditel'nosti detektorov granic [On some issues of the quantitative performance evaluation of edge detectors]. *Programmnie produkti i sistemi*. No 4. P. 13–20.

[12] Osipov A. S. 2016. Ob ispol'zovanii elementov nechetkoi logiki v ocenke algoritmov identifikacii lic [On the use of fuzzy logic in evaluation of the face detection algorithms]. *Moscow SRISA RAS Publ.* 6(2):62–69.

[13] Osipov A. S. 2017. Nechetkie meri i ih ispol'zovanie v ocenke algoritmov komp'uternogo zreniya [On the use of fuzzy similarity measures in evaluation of the computer vision algorithms]. *Moscow SRISA RAS Publ.*, 2017. 7(1):46–57.

[14] The ORL database of faces AT&T laboratories Cambridge. URL: `http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html` (accessed October 30, 2017).

[15] Levin B. A., A. S. Bugaev, S. I. Ivashov, and V. V. Razevig. 2013. Distancionno-pilotiruemie apparati i bezopasnost' puti [Distantly piloted aircrafts and the track security]. *Mir Transporta*. 11(2):152–157.

[16] Johnson C. 2013. Image processing techniques for the detection and characterisation of features and defects in railway tracks // PhD Thesis, Manchester Metropolitan University, UK.

Received May 28, 2018