СЕНСОРНЫЕ СИСТЕМЫ, 2019, том 33, № 4, с. 287–304

ОБЗОРЫ

УДК :159.938.25 + 004.05

СУБЪЕКТИВНАЯ ОЦЕНКА КАЧЕСТВА СТАТИЧЕСКИХ И ВИДЕОИЗОБРАЖЕНИЙ: МЕТОДОЛОГИЧЕСКИЙ ОБЗОР

© 2019 г. М. А. Грачева[1,*], В. П. Божкова[1], А. А. Казакова[1,2], Г. И. Рожкова[1]

[1] Институт проблем передачи информации им А.А. Харкевича РАН, 127051 Москва, Б. Каретный пер., 19, стр. 1, Россия

[2] Российский национальный исследовательский медицинский университет им. Н.И. Пирогова Минздрава России, 117997 Москва, ул. Островитянова, д. 1, Россия

**E-mail: mg.iitp@gmail.com*

Поступила в редакцию 10.04.2019 г.
После доработки 17.06.2019 г.
Принята к публикации 22.07.2019 г.

Для сравнения алгоритмов сжатия, фильтрации или трансформации визуальных данных, разрабатываемых с целью оптимизации систем хранения, передачи и анализа информации, используются различные методы оценки качества получаемых изображений. Данные методы делятся на объективные – основанные на четко сформулированных математических критериях, не зависящих от мнения людей, и субъективные – основанные на привлечении людей и анализе их мнений. В настоящем обзоре дано краткое представление о методологических аспектах исследований, посвященных субъективным методам оценки качества изображений. В настоящее время такие исследования являются очень востребованными, а в отечественной литературе нет руководств для практического применения этих методов разработчиками, не имеющими опыта психофизиологических исследований. Рассмотрены наиболее часто используемые методы оценки и сравнения статических и видеоизображений: ACR – абсолютный категориальный рейтинг, ACR-HR – абсолютный категориальный рейтинг со скрытым эталоном, SSCQE – непрерывная оценка качества одиночных стимулов, DCR – категориальная оценка ухудшений, DSCQR – оценка на основе парных стимулов, PC – попарное сравнение, PSJ – попарная оценка сходства, SDSCE – непрерывная оценка качества на основе одновременных сдвоенных стимулов. Приведены также общие рекомендации по планированию и проведению экспериментов с привлечением людей.

Ключевые слова: качество изображений, субъективная оценка, исследования на пользователях, дизайн эксперимента

DOI: 10.1134/S0235009219040036

ВВЕДЕНИЕ

В современных технических системах передачи, хранения и воспроизведения визуальной информации широко применяются различные способы сжатия и фильтрации изображений с целью повышения скорости передачи и эффективности использования носителей информации, комфортности восприятия и удобства анализа визуальных данных.

Для сравнения разрабатываемых алгоритмов обработки фото или видео используются различные методы оценки качества получаемого изображения. Методы оценки делятся на два типа “объективные (количественные)” – использующие алгоритмы оценки качества по набору формализованных критериев, и “субъективные (экспертные)” – основанные на привлечении людей и анализе их мнений (Phillips, Eliasson, 2018). Методы второго типа – субъективные методы оценки – фигурируют в литературе также под терминами “исследования на людях” (human studies), “исследования на пользователях” (user studies) и “пользовательские тестирования” (usability testings).

Согласно буквальному смыслу терминов, объективные оценки – это оценки по четко сформулированным критериям, не зависящие от мнения людей и адекватно отражающие оцениваемое качество, а субъективные оценки – это оценки по индивидуальным и часто неосознаваемым (“скрытым”) критериям, в связи с чем они могут сильно варьироваться. В данном обзоре обращается внимание на то, что, в принципе, субъективные и объективные оценки должны хорошо соответствовать друг другу, для чего существующие методы обоих типов нужно целенаправленно совершенствовать. Что касается объективных методов, то, когда потребителем контента является че-

ловек, используемые при автоматизированной обработке критерии нужно приближать к “человеческим”, а что касается субъективных методов, то нужно оптимизировать организацию проводимых исследований и экспериментальные парадигмы, устраняя факторы, приводящие к большой вариабельности и возможной неадекватности оценок. Очевидно, что для получения воспроизводимого результата общая оценка должна быть основана на мнениях большого числа разных индивидуумов; поэтому при правильной организации исследования субъективный метод должен дать результат, который по своей точности и повторяемости не уступает объективному методу.

Объективная оценка в традиционном смысле предполагает использование метрик оценки качества изображения. Использование критериев качества в настоящий момент является необходимым атрибутом хорошей научной работы в области обработки изображений. Формальные математические критерии, основанные на физических параметрах изображения, имеют ряд очевидных достоинств (Монич, Старовойтов, 2008; Phillips, Eliasson, 2018). В идеале объективные методы могут стать основой полной автоматизации анализа изображений, что сделает оценку визуального контента значительно менее ресурсоемкой.

Если в рассматриваемой задаче для входных данных известно идеальное изображение, то критерии качества обычно строятся на основе метрики различий двух ответов алгоритма на идеальное и реальное изображения. На практике же во многих работах для этого используются легко интерпретируемые, но никак не обоснованные метрики (как, например, СКО – среднее квадратичное отклонение, в англ. MSE – Mean-Squared Error) (Chen et al., 2017; Ledig et al., 2017). Часто встречается также комбинация из трех метрик: СКО (MSE), пиковое отношение сигнала к шуму (в англ. PSNR – Peak Signal-to-Noise Ratio) и индекс структурного сходства (в англ. SSIM – Structure Similarity) (Lu et al., 2015; Chen et al., 2017; Anwar et al., 2018).

В некоторых случаях, когда идеальный стимул неизвестен, используются неэталонные оценки, которые основываются только на текущем изображении. Как правило, такие метрики используют сложный математический и статистический аппарат и являются узкоспециализированными, применяемыми только для определенного вида искажений, например в процессе подводной съемки. Ли и соавт. (Li et al., 2019) описали явления, сопровождающие распространение света в воде и вызываемые ими искажения изображений, для компенсации которых необходима последующая компьютерная обработка. Комплекс этих искажений включает: затуманивание изображений (снижение контрастности), трансформацию цветов, зашумление, неоднородность освещения фрагментов сцены и виньетирование. Для улучшения качества подводных изображений предложены разные методы, эффективность которых нужно оценить.

В частности, в работе Мангеруга и соавт. (Mangeruga et al., 2018, a) для неэталонной оценки улучшения качества подводных фотографий использовались такие метрики, как средняя яркость изображения, информационная энтропия и средний цветовой градиент, применимость которых для оценки качества изображений была обоснована в другой работе (Xie, Wang, 2010). Не будучи в них полностью уверенными, авторы применили также и субъективную методологию в виде оценки группой экспертов в области восстановленных подводных изображений (Mangeruga et al., 2018, b). Эта группа проводила оценку качества улучшенных изображений на основании опросника. Опросник был составлен в виде мозаики картинок, включающей исходное изображение и изображения, обработанные разными алгоритмами. К каждой мозаике прилагалась таблица множественного выбора, в которой эксперт проставлял оценки от единицы до пяти для всех изображений, принимая во внимание результат как цветокоррекции, так и повышения контраста. Оценки, данные каждым экспертом по каждому изображению, выбранному из большого датасета, представляли собой совокупность данных, которая затем интерпретировалась. Один из возможных способов извлечения полезной информации из экспертной оценки – подсчитать средний результат голосования экспертов по каждому алгоритму с определением достоверности различий.

Основным недостатком упомянутых математических объективных метрик является их слабая корреляция с субъективными оценками. Изображение, высоко оцененное при помощи субъективного метода, может иметь низкую оценку качества по какой-либо объективной метрике, и наоборот. Поэтому наряду с такими объективными метриками стали разрабатываться и метрики, обоснованные на уровне неких общих представлений о зрительной системе человека (ЗСЧ) (Wang et al., 2004; Lissner et al., 2013). Они необходимы, в частности, для задач обработки изображений, предназначенных для визуализации, где логично не учитывать различия, игнорируемые ЗСЧ. Например, развиваются модели, в которых принимаются во внимание конкретные свойства пространственно-частотной функции контрастной чувствительности ЗСЧ (Божкова и др., 2019). Эти свойства заложены в такие метрики качества, как S-CIELAB и iCAM (Zhang, Wandell, 1996; 1998; Fairchild, Johnson, 2004; Kuang et al., 2007). Более подробно они обсуждены Филлипсом и

Элайсоном (Phillips, Eliasson, 2018). Отдельной трудной и практически неисследованной проблемой является учет индивидуальных особенностей ЗСЧ. При этом индивидуальная вариативность ЗСЧ в области контрастной чувствительности (Kim et al., 2017), аномалий цветовосприятия (Максимов, 1984) и остроты зрения (Рожкова, Матвеев, 2007) изучена достаточно подробно. Активно разрабатываются математические модели функционирования первых уровней зрительной системы человека – сетчатки и первичных зон зрительной коры (Лебедев, 2015; Watson, Ahumada, 2005; 2008; 2012; Watson et al., 2009; Kontsevich, Tyler, 1994; 2013). Следует заметить, что построение полной, подробной и точной модели восприятия изображений зрительной системой человека невозможно в принципе (всегда найдутся люди, не укладывающиеся в рамки данной модели), а проблема построения адекватных метрик качества изображений стоит достаточно остро, особенно в области сжатия видеопотоков (Ватолин, Паршин, 2006; Боков, Ватолин, 2016).

Со временем, по-видимому, будут созданы модели ЗСЧ, предназначенные для разного контингента и разных задач, и свойства этих моделей будут учтены в соответствующих объективных метриках. Однако на данном этапе использование экспертного подхода остается самым адекватным и востребованным методом оценки качества изображений, несмотря на то, что применение этого метода требует четкой организации процесса тестирования и большого числа экспертов, больших затрат времени и материальных ресурсов.

Чтобы получить корректные субъективные оценки изображений с привлечением испытуемых, необходимо соблюдать основные правила планирования экспериментов, а для сопоставления оценок, полученных разными исследователями в разных экспериментах, необходимо еще и соблюдать одинаковый протокол эксперимента. Для стандартизации проводимых разными исследователями экспериментов с привлечением людей международная организация, называемая в России Международным Союзом Электросвязи (МСЭ), а за рубежом – International Telecommunication Union (ITU) – разрабатывает соответствующие рекомендации.

В данном обзоре рассмотрены основные методы субъективной оценки статических и видеоизображений, рекомендуемые МСЭ (ITU). Большинство рекомендаций МСЭ (ITU) опубликованы в открытом доступе как на английском, так и на русском языках. Специалисты продолжают исследовать особенности существующих методов субъективной оценки, проводить сравнительные работы и публиковать дополнения к уже описанным методам, поэтому рекомендации МСЭ (ITU) обновляются.

Далее в тексте будут представлены описания методов с аббревиатурами их названий. Данные аббревиатуры введены МСЭ (ITU), а не авторами. Использование аббревиатур может несколько затруднять чтение текста, но все эти термины являются общепринятыми и широко используются в публикациях по теме, поэтому представляется нерациональным и некорректным отказываться от их использования в обзоре.

Существуют два вида экспертных оценок: абсолютные и сравнительные. В первом случае наблюдатель вынужден принимать решение только на основании своего собственного опыта и должен оценить качество изображения по какой-то заранее определенной шкале. Иногда процесс оценивания облегчается тем, что наблюдателю предлагается набор эталонных изображений. При сравнительных оценках наблюдатель должен ранжировать набор конкретных изображений, т.е. расставить их в ряд по возрастанию/убыванию качества.

Наиболее часто используемые методы оценки и сравнения статических и видеоизображений (Mantiuk et al., 2012; Xu et al., 2015; ITU-R P.910, 1999) “с одним стимулом” – **single stimulus (SS):** ACR (absolute category rating), ACR-HR (absolute category rating with hidden reference), SSCQE (single stimulus continuous quality estimation); “с двумя стимулами” – **double stimulus (DS):** DCR (degradation category rating), DSCQR (double stimulus continuous quality rating), PC (pair comparison), PSJ (pairwise similarity judgement), SDSCE (simultaneous double stimulus for continuous evaluation). Далее по тексту даны расшифровки и подробные описания методов.

Иногда методы делят на методы с явными эталонами (идеальными изображениями) – “explicit reference” (например, метод DCR) и те, в которых явные эталоны не используются (методы ACR, ACR-HR, PC). Очевидно, что только методы с явными эталонами позволяют оценить прозрачность или точность передачи (обработку называют *прозрачной* при отсутствии видимой разницы между обработанным стимулом и исходным эталоном). Многие методы предполагают возможность последовательного (“sequential”) или одновременного (“simultaneous”) предъявления стимулов. Для методов, в которых эти режимы возможны, на схемах представлены оба варианта.

В разделах, описывающих каждый из методов, указаны их альтернативные названия, встречающиеся в литературе. Всем описанным методам присущи как достоинства, так и недостатки (Macdiarmid, Darby, 1982; Narita, Sugiura, 1997; Corriveau et al., 1999; Mantiuk et al., 2012), поэтому выбор подходящего метода и его адаптация под

задачи работы всегда остаются на усмотрение исследователя.

МЕТОДЫ, ОСНОВАННЫЕ НА ПРЕДЪЯВЛЕНИИ ОДИНОЧНЫХ СТИМУЛОВ

Абсолютный категориальный рейтинг (ACR: Absolute category rating)

Базовый метод оценки качества одиночных изображений, далее часто называемых стимулами (как принято в психофизиологических исследованиях), – абсолютный категориальный рейтинг (ACR). В литературе этот метод также встречается под названиями “показатель абсолютной категории”, “single stimulus category scale/rating”, “SSQS” или “SSQR”.

Метод ACR предполагает единовременную оценку одного стимула: тестовые изображения демонстрируются поодиночке и независимо оцениваются по шкале категорий. После предъявления и оценки первого стимула предъявляется следующий, и так до конца последовательности (рис. 1).

Рекомендуемая шкала для оценки – от 1 до 5 (ITU-R P.910, 1999), со следующими условными обозначениями категорий: 1 – “плохо”, 2 – “неудовлетворительно”, 3 – “удовлетворительно”, 4 – “хорошо”, 5 – “отлично”. Однако используются и иные шкалы, например, с оценками от 1 до 9, от 0 до 10 и другими (подробнее об этом сказано далее). Время выбора ответа (голосования) предлагается ограничивать десятью секундами. Для анализа результатов рекомендуется использовать среднее значение оценки с указанием стандартного отклонения.

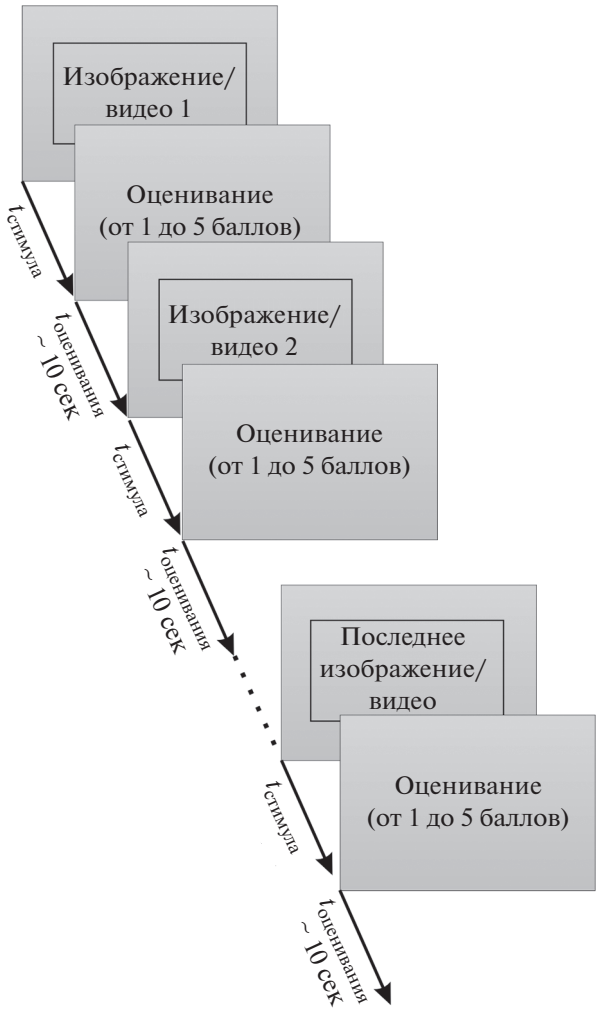

Рис. 1. Схема метода ACR (абсолютный категориальный рейтинг).

Абсолютный категориальный рейтинг со скрытым эталоном (ACR-HR: Absolute category rating with hidden reference)

Метод ACR-HR (рис. 2), абсолютный категориальный рейтинг, также встречается в литературе под названием “показатель абсолютной категории со скрытым эталоном”. Данный метод похож на метод ACR: тестовые статические или видеоизображения демонстрируются поодиночке и независимо оцениваются по шкале категорий; после предъявления и оценки первого стимула предъявляется второй, второй стимул оценивается, и так до конца последовательности. Отличие метода ACR-HR от обычного ACR в том, что помимо обработанных изображений, качество которых требуется оценить, в последовательность стимулов включены необработанные образцовые (референтные) стимулы – эталоны (reference stimuli). Испытуемые не знают, какие из стимулов эталонные и оценивают их наравне с остальными стимулами. Как и в методе ACR, рекомендуемое время для ответа (голосования) – 10 с.

Итоговая метрика основана на разности оценок эталонов и обработанных изображений. Показано, что оценки испытуемых могут существенно различаться для разных сцен (в зависимости от содержания сцены) и могут зависеть от других параметров стимула, а не только от их качества (van Dijk et al., 1995). Метрика, основанная на включении эталонов, позволяет отнормировать оценки относительно субъективного восприятия качества.

Согласно (ITU-T P.930, 1996), для итоговой оценки рекомендуется использовать следующую формулу:

$$D = R_{\text{stimuli}} - R_{\text{reference}} + 5,$$

где D – разностная оценка, R_{stimuli} – рейтинг тестовых изображений, $R_{\text{reference}}$ – рейтинг эталонных изображений.

Если обработанный стимул получает более высокую оценку, чем эталонный, то $D > 5$. Согласно рекомендациям МСЭ (ITU), такие оценки считаются действительными (хотя не уточнено, приравниваются ли такие баллы к оценке “отлично” или им дают другую интерпретацию). В некоторых случаях, когда такое превышение нежелательно, рекомендуется использовать следующую формулу:

$$D_{\text{crushed}} = (7*D)/(2 + D), \quad \text{когда} \quad D > 5.$$

Не все авторы используют для дифференциальной оценки именно эту, рекомендованную МСЭ (ITU), формулу. Например, в работе (Mantiuk et al., 2012) авторы используют упрощенную формулу, без прибавления константы 5.

Те же авторы предлагают учитывать, что разные испытуемые склонны по-разному воспринимать привязку к используемым уровням шкалы, и советуют использовать для унификации ответов z-преобразование для приведения оценок каждого пользователя к общему среднему и общему стандартному отклонению. При большом числе испытуемых можно использовать также другие преобразования (Torgerson, 1985).

Корректная оценка качества обработки может быть проведена только для тех изображений, которым в эталонном виде эксперты дают оценку “хорошо” или “отлично” (иначе шкала оценок обрезается сверху).

Ограничением метода ACR-HR является необходимость иметь эталонные стимулы, которые должны не слишком явно отличаться от тестовых (только по качеству после обработки). Например, метод ACR-HR хорошо подходит для оценки алгоритмов сжатия, когда в одной последовательности могут быть как оригинальные изображения, так и обработанные несколькими выбранными алгоритмами. Однако этот метод может не подойти, например, при оценке алгоритмов ретаргетинга, когда оригинальные изображения зачастую отличаются от обработанных по соотношению сторон, что делает эталон выделяющимся для испытуемого и снижает ценность предъявления эталонов в ряду обработанных изображений.

Рис. 2. Схема метода ACR-HR (абсолютный категориальный рейтинг со скрытым эталоном).

Непрерывная оценка качества одиночных стимулов (SSCQE: Single Stimulus Continuous Quality Estimation)

Метод непрерывной оценки качества одиночных стимулов, SSCQE, схематически представлен на рис. 3. В литературе метод SSCQE встречается также под аббревиатурой “SSCQR” – “single stimulus continuous quality rating”, и в переводной версии под названием “метод непрерывной оценки качества с одним источником воздействия”. В отличие от предыдущих методов – ACR и ACR-HR, – которые могут использоваться как для оценки изображений, так и для оценки видео, метод SSCQE предназначен только для оценки видео.

Метод можно реализовать, имея ручное записывающее устройство с ползунковым механизмом; рекомендованный диапазон перемещения ползунка – около 10 см.

Испытуемые просматривают оцениваемый материал только 1 раз; эталонное видео им не предъявляют. В течение просмотра испытуемый указывает оценку качества видео в каждый момент, двигая ползунок. Шкала ползунка соответствует шкале непрерывной оценки качества (с пятью условными вспомогательными категория-

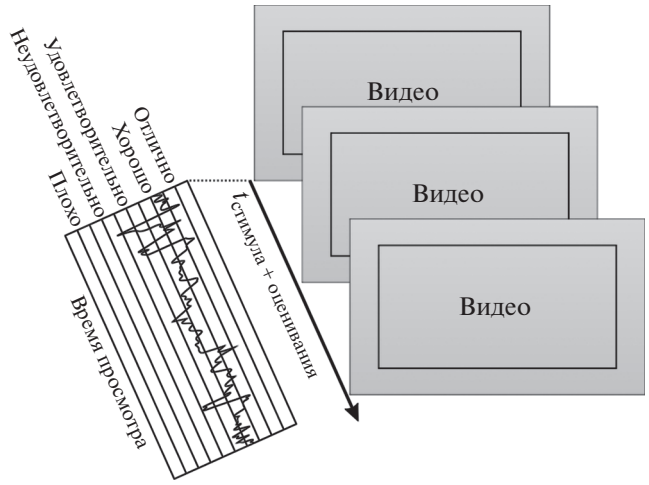

Рис. 3. Схема метода SSCQE (непрерывная оценка качества одиночных стимулов).

ми: “отлично”, “хорошо”, “удовлетворительно”, “плохо”, “неприемлемо”).

Фрагменты видео, подвергнутые одному из сравниваемых способов обработки изображений, по длительности должны быть не менее пяти минут (без перерыва); общая рекомендуемая разными авторами длительность сеанса – 30–60 мин. В течение сеанса фрагменты демонстрируются без разделения, при этом все варианты видеофрагментов и все варианты обработки должны хотя бы раз быть предъявлены в течение одного сеанса (т.е. всего может быть не более двенадцати сравниваемых типов обработки). Метод рассчитан именно на использование длительных тестовых отрезков и максимально приближен к реальным условиям просмотра видео (поэтому и отсутствуют эталонные стимулы). Все сочетания видеофрагментов/условий обработки должны быть предъявлены одинаковому числу участников (но не обязательно одним и тем же) (ITU-T BT.500-13, 2002; Alpert, Evian, 1997; Kratochvíl, Slanina, 2010).

В работе (Pinson, Wolf, 2003) авторы предлагают модифицировать метод, введя в него скрытый эталон (т.е. испытуемый должен оценивать эталон как обычное видео, не зная, какое именно видео – эталонное), и потом оценки эталонного видео вычитаются из оценок измененного (тестового) видеоряда (“hidden reference removal”).

Развитием метода SSCQE с введением постоянно присутствующего на соседнем экране (или половине одного экрана) эталонного стимула является метод SDSCE[1] (метод непрерывной оценки качества на основе одновременных сдвоенных стимулов).

Об оценке данных, получаемых в результате использования данного метода, подробно написано в рекомендациях (ITU-T BT.500-13, 2002).

Существуют свидетельства того, что при пользовании методами моментальной оценки (SSCQE, SDSCE) испытуемые склонны несимметрично оценивать ухудшения и улучшения видео: снижение качества оценивается большинством пользователей быстро, а улучшение – с задержкой (Hamberg, Ridder, 1999). Эту асимметрию важно принимать во внимание при анализе результатов.

Метод SSCQE был предложен в 1995 г. Хамбергом и де Ридером для 2D видеопоследовательностей (Hamberg, de Ridder, 1995), а несколькими годами позже Айзелштейн и соавт. (Ijsselsteijn et al., 1998) применили его для стереоизображений. Впоследствии методы с “одним стимулом” многократно использовались для оценивания 2D и 3D контента (Aldridge et al., 1998; Yano et al., 2002; Redi et al., 2010; Lambooij et al., 2011; Mantiuk et al., 2012).

МЕТОДЫ, ОСНОВАННЫЕ НА ПРЕДЪЯВЛЕНИИ ПАР СТИМУЛОВ

Многие методы предполагают предъявление сдвоенных стимулов, которые либо оцениваются одновременно, либо сравниваются. При этом в части методов в качестве второго (парного) стимула используется эталон, в части – другой стимул, а в некоторых может попеременно использоваться эталон или тестовый стимул (т.е. в данных случаях эталон всегда “скрытый”).

Категоризация методов по типу парного стимула:

– явный эталон: DCR по рекомендациям (ITU-R P.910, 1999) (degradation category scale), SDSCE (simultaneous double stimulus for continuous evaluation);

– скрытый эталон: DCR по работе (Mantiuk et al., 2012) (degradation category scale), DSCQS (double stimulus continuous quality scale), PC (pair comparison, вариант со скрытым эталоном), PSJ (pairwise similarity judgement, вариант со скрытым эталоном);

– другой тестовый стимул: PC (pair comparison, вариант с разными тестовыми стимулами), PSJ (pairwise similarity judgement, вариант с разными тестовыми стимулами).

Категориальная оценка ухудшения (DCR: Degradation category rating)

Два типа организации экспериментов по методу DCR (категориальная оценка ухудшений) представлены на рис. 4. Аббревиатура DCR ис-

[1] По неизвестным причинам в некоторых источниках метод SDSCE называют DSCQS, хотя они очень сильно различаются. Мы используем терминологию, введенную стандартами (ITU-T BT.500-13,2002), но в некоторых местах добавляем сноски о двойственности этих аббревиатур в литературе.

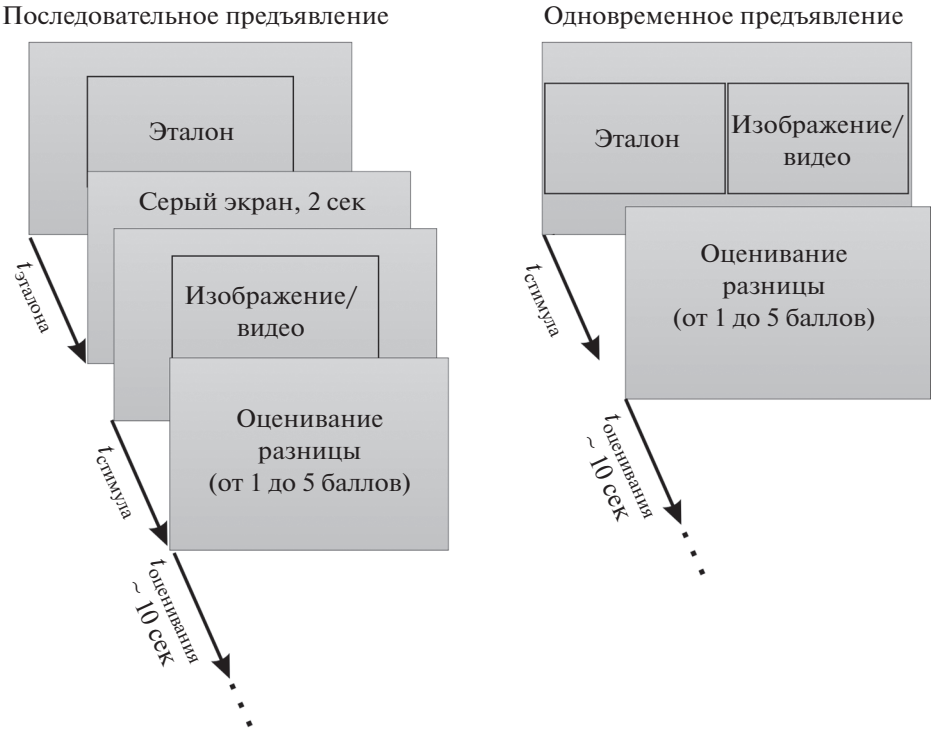

Рис. 4. Схема метода DCR (категориальная оценка ухудшения качества).

пользуется в (ITU-R P.910, 1999). Другие встречаемые в литературе аббревиатуры и названия метода: “double stimulus impairment scale” – “DSIS[2]”, “double stimulus comparison scale” – “DSCS[3]”, “оценка ухудшения категории качества”, “шкала ухудшения с двумя источниками воздействия”.

При использовании метода DCR стимулы предъявляются парами: эталон и тестовый стимул, при этом возможно последовательное или одновременное предъявление. В первом случае сначала предъявляется эталон[4], и между последовательно предъявляемыми стимулами в течение двух секунд предъявляется серый экран (рис. 4, левый ряд).

После предъявления обоих стимулов испытуемые должны оценить степень ухудшения качества тестового изображения относительно эталона по 5-балльной шкале (5 – “незаметное”, 4 – “заметное, но не раздражающее”, 3 – “немного раздражающее”, 2 – “раздражающее”, 1 – “сильно раздражающее”).

Рекомендуемое время для голосования, как и в методах ACR и ACR-HR – 10 с (ITU-R P.910, 1999; ITU-T BT.500-13, 2002).

Метод оценки на основе парных стимулов с использованием непрерывной шкалы

(DSCQS[5]: double stimulus continuous quality scale)

В схематическом виде метод оценки на основе парных стимулов с использованием непрерывной шкалы представлен на рис. 5. Другие аббревиатуры и названия, встречающиеся для этого метода в литературе: “double stimulus continuous quality rating/estimation”, “DSCQR” или “DSCQE”, “метод двух источников воздействия с непрерывной шкалой качества”.

При реализации данного метода испытуемому предъявляются два стимула: тестовый и эталонный. Испытуемый не знает, какой из стимулов – эталон, и должен оценить оба изображения по непрерывной шкале, ставя соответствующие метки. Для упрощения использования непрерывной

[2] Название DSIS используется в (ITU-T BT.500-13, 2002).

[3] В работе (Pinson, Wolf, 2003) встречается аббревиатура DSCS. Остается неизвестным, почему авторы используют именно ее, т.к. больше она нигде не встречается.

[4] В рекомендациях (ITU-R P.910, 1999) указано, что эталон всегда должен предъявляться первым и что испытуемый всегда должен знать, какой из стимулов – эталонный, а какой – тестовый. Однако в работе (Mantiuk et al., 2012) под тем же названием (DCR) рассматривается метод, при котором тестовый и эталонный стимулы предъявляются в случайном порядке и после предъявления оба стимула оцениваются по шкале 1-5.

[5] Термин DSCQS для такого алгоритма предъявления используется в (ITU-T BT.500-13, 2002). В работе (Alpern, Evian, 1997) под аббревиатурой DSCQE фигурирует метод, по описанию идентичный SDSCE, что некорректно.

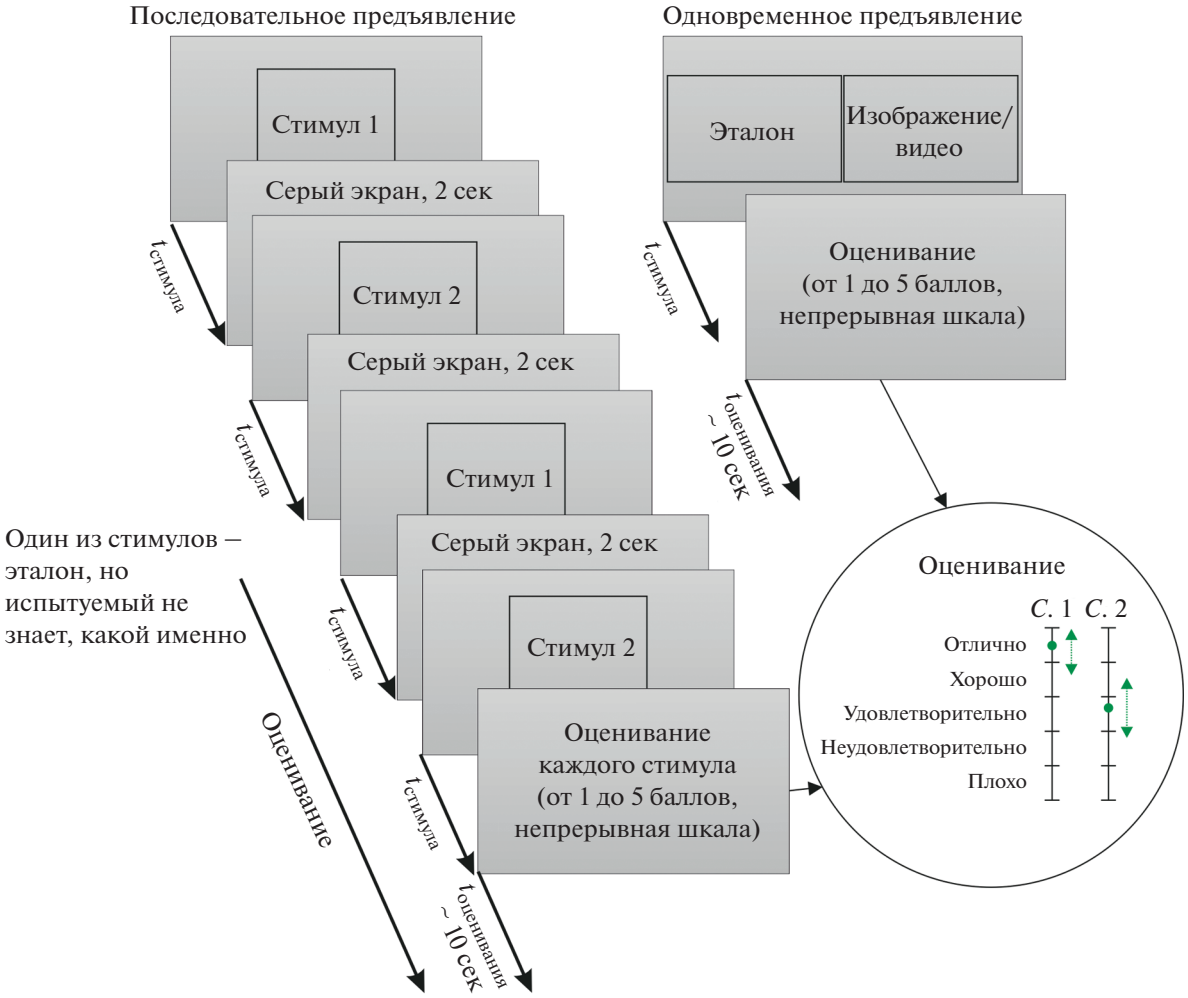

Рис. 5. Схема метода DSCQS (метод оценки на основе парных стимулов с использованием непрерывной шкалы).

шкалы на ней расставляются “якоря” категорий (“отлично”, “хорошо”) и другие.

Существуют два основных варианта реализации метода: последовательный и одновременный. В первом случае стимулы показываются по очереди 2 раза (т.е. демонстрируется первый стимул, затем второй, снова первый и снова второй); между стимулами в течение двух секунд демонстрируется серый экран. Во время второго предъявления испытуемый может уже ставить оценки обоим стимулам. После всей последовательности предъявления испытуемому дается дополнительное время для голосования (около десяти секунд). Рекомендуемая длительность предъявления каждого стимула[6] – тоже 10 с (ITU-T BT.500-13, 2002). В случае одновременного предъявления оба стимула демонстрируются рядом (на двух половинах дисплея или на соседних экранах) в течение десяти секунд, затем испытуемому дается 10 с для голосования.

Полученные оценки для тестового и эталонного стимулов переводятся в шкалу от 0 до 100, вычисляется разность между двумя оценками. Подробную информацию об анализе полученных данных можно найти в (ITU-T BT.500-13, 2002).

Отмечается, что оценки, полученные по методу DSCQS, часто ошибочно связывают с использованными при тестировании категориями (“отлично”, “хорошо”, “удовлетворительно”), но такое сопоставление баллов с категорией некорректно.

В ряде работ обсуждалось, какие методы (с эталоном или без) следует применять для получения наиболее точных оценок качества изображения. В работе (Macdiarmid, Darby, 1982) авторы пытались доказать, что DSCQE дает более точные значения благодаря наличию эталона. Такая точка зрения была оспорена другими исследователями (Narita, 1994; Narita, Sugiura, 1997), которые

[6] См. также комментарий про палиндромное предъявление в разделе “Длительности предъявления стимулов”.

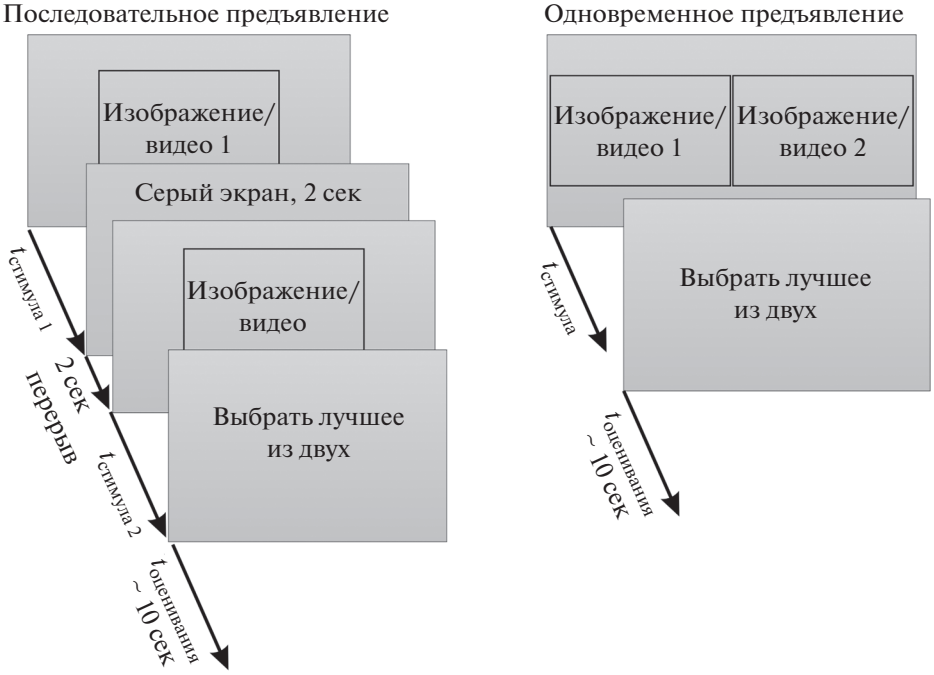

Рис. 6. Схема метода PC (попарное сравнение).

высказались за использование ACR – метода без эталона с дискретной шкалой от 1 до 5. Эта давняя дискуссия пока ничем не разрешилась, и все методы продолжают использоваться при субъективном тестировании (Stoica et al., 2003; Hands, 2004; Pinson et al., 2015; Phillips, Eliasson, 2018).

Попарное сравнение

(PC: Pair comparison)

Метод попарного сравнения, PC, также встречается в литературе под названиями “pairwise comparison”, “two-alternative forced choice (2AFC)” в случае одновременного предъявления и “two-interval forced choice (2IFC)” – в случае последовательного.

Стимулы предъявляются парами (одновременно или последовательно; в случае последовательного предъявления между двумя стимулами в течение двух секунд предъявляется серый экран) (рис. 6). Стимулы представляют собой обработанный двумя разными способами эталонный стимул. Испытуемый обязательно должен выбрать одно изображение как лучшее, даже если не видит разницы между ними (“процедура вынужденного выбора” или “forced-choice procedure”). Время голосования, как и в предыдущих методах, рекомендуется ограничить десятью секундами.

Согласно рекомендациям (ITU-R P.910, 1999), следует предъявлять n(n-1) возможных комбинаций, т.е. все пары предъявляются по 2 раза в разном порядке (это особенно важно при последовательном предъявлении). Согласно работе (Mantiuk et al., 2012), можно сократить число комбинаций до n(n-1)/2, т.е. каждую пару предъявлять только 1 раз. Есть также другие варианты сокращения процедуры, например, за счет использования сбалансированных неполных блок-схем (balanced incomplete block design) (Хикс, 1967; Gulliksen, Tucker, 1961) или за счет использования сортировочного алгоритма для выбора пар (Silverstein, Farrell, 2001). В случае алгоритмов с сокращенным числом пар обычно принимается условие: если стимул А оценен ниже, чем стимул Б, а стимул В – выше, чем стимул Б, то стимул В считается оцененным выше, чем стимул А (что не всегда соответствует реальности).

Иногда используется также и метод сравнения по тройкам стимулов (“three-alternative forced choice – 3AFC”), описанный, например, в (Phillips, Eliasson 2018), но в рекомендациях МСЭ (ITU) такой метод не встречается.

Попарная оценка сходства

(PSJ: pairwise similarity judgement)

Метод попарной оценки сходства, PSJ, похожий на DCR и PC, подробно описанный в рекомендациях МСЭ (ITU-T BT.500-13, 2002; ITU-T P.800, 1996), рассматривается как один из основных методов в книге (Xu et al., 2015) и работах (van Dijk et al., 1995; Mantiuk et al., 2012; Mohammadi et al., 2014). Схематически метод представлен на рис. 7.

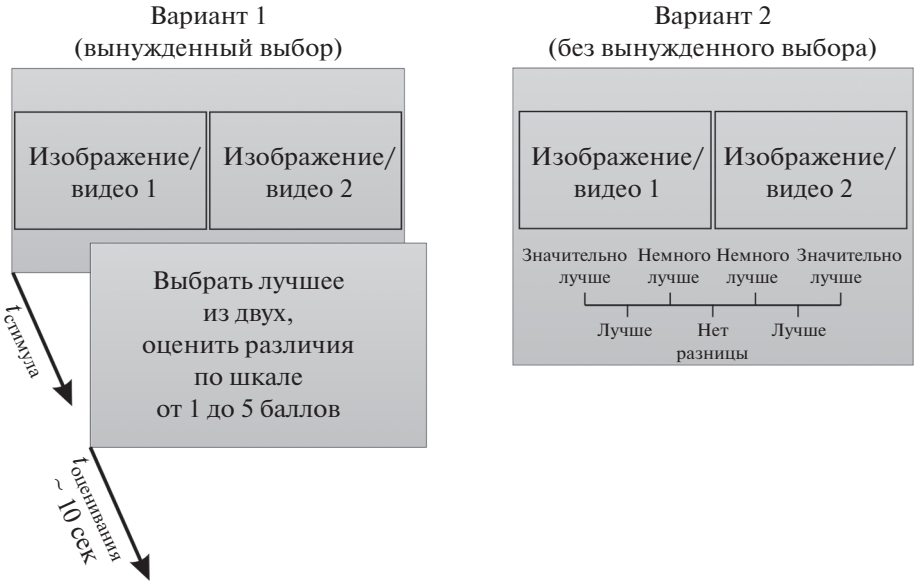

Рис. 7. Схема метода PSJ (попарная оценка сходства).

Стимулы предъявляются парами, и испытуемый должен выбрать лучший из стимулов (даже если не видит разницы), а затем указать по шкале от 1 до 5 степень различия между изображениями (Xu et al., 2015). Другой разновидностью алгоритма из работы (Mantiuk et al., 2012) является вариант, когда при предъявлении стимула испытуемый оценивает различия сразу по 7-уровневой шкале от –3 до 3 (0 – "нет разницы", 1 – "немного лучше", 2 – "лучше", 3 – "значительно лучше" для каждого изображения; при выборе левого изображения в качестве лучшего различиям присваивается один знак, при выборе правого – другой). В данном случае процедура уже не является процедурой вынужденного выбора (forced choice), так как испытуемый может выбрать 0 – отсутствие ощутимых различий.

Число пар для исследования выбирается так же, как и при обычном PC-методе (сортировочный алгоритм с оценкой "0" присваивает изображениям ранги случайным образом).

Анализ получаемых данных рассмотрен в рекомендациях МСЭ (ITU) и в работе (Mantiuk et al., 2012).

Непрерывная оценка качества на основе одновременных сдвоенных стимулов

(SDSCE: simultaneous double stimulus for continuous evaluation)

Метод непрерывной оценки качества на основе одновременных сдвоенных стимулов, SDSCE[7], в рекомендациях МСЭ (ITU) также встречается под русскоязычным названием "метод с двумя источниками одновременного воздействия для непрерывной оценки".

Для реализации метода рекомендуется использовать ручное записывающее устройство с ползунковым механизмом, как и в методе SSCQE. Схематически метод представлен на рис. 8.

SDSCE является развитием метода SSCQE, но использует не одно видео, а два: тестовое и эталонное, показываемые одновременно (на двух экранах или каждый на половине экрана). Испытуемый во время просмотра сравнивает эталонное и тестовое видео и двигает ползунок устрой-

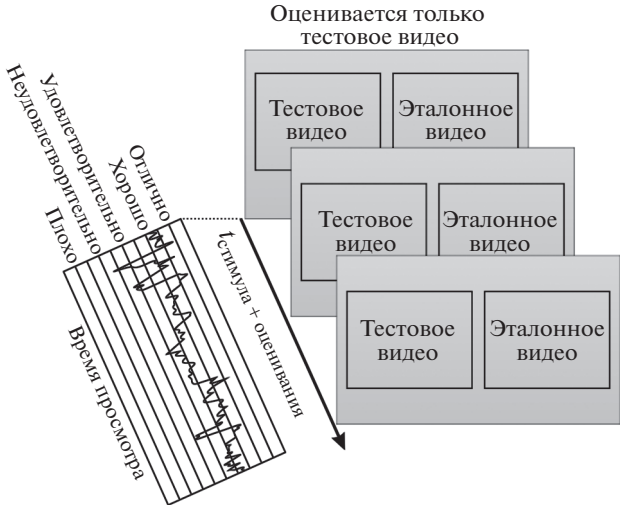

Рис. 8. Схема метода SDSCE (непрерывная оценка качества на основе одновременных двойных стимулов).

[7] В работе (Alpern, Evian, 1997) аналогичный по описанию метод обозначен аббревиатурой DSCQE, но такое использование аббревиатуры некорректно.

ства, указывая моментальную оценку качества (ITU-T BT.500-13, 2002).

В данном методе важно проводить этап обучения и демонстрационный сеанс (оценки которых не учитываются). После обучения проводится проверка того, что при демонстрации пары эталонных видео рядом (и в качестве “эталона”, и в качестве “тестового”), испытуемые не видят отличий – оценка сходства близка к максимальной. Если это условие не выполняется, необходимо проверить демонстрационную систему, повторить инструктирование и повторно провести проверку. Если повторный инструктаж не помогает, приходится исключать испытуемого из выборки.

Преимущества метода: возможность моментальной оценки качества и возможность демонстрации длительных видео, что ближе к реальным условиям просмотра.

Считается, что данный метод ориентирован скорее на оценку точности передачи, а не качества самого изображения (так как тестовое видео все время сравнивается с эталоном) (Alpern, Evian, 1997).

Недостатками метода считаются необходимость перевода взгляда (и внимания) с одного стимула на другой и разная скорость реакции испытуемых. Также нужно помнить, что испытуемый может терять ощущение абсолютного положения ползунка на шкале (переводить его выше или ниже предыдущей оценки, но не соотносить выставляемую оценку с крайними положениями шкалы) (Pinson, Wolf, 2003).

ОБЩИЕ РЕКОМЕНДАЦИИ ПО ПРОВЕДЕНИЮ ЭКСПЕРИМЕНТА

Инструктаж. Перед началом испытаний обязательно проводится инструктирование испытуемых. Инструкции рекомендуется выдавать участникам в письменном виде, чтобы убедиться, что все испытуемые получили одинаковую инструкцию (дополнительно ее нужно дублировать устно), на все вопросы испытуемых нужно дать по-возможности исчерпывающие ответы.

Обучающие стимулы. Для снижения эффекта обучения в процессе тестирования и лучшего понимания задачи испытуемым рекомендуется включать обучающие последовательности в начало эксперимента. Оценки, полученные на обучающих последовательностях, не учитываются в итоговых результатах. В рекомендациях (ITU-T BT.500-13, 2002) для привыкания испытуемых к условиям тестирования в схему эксперимента предлагается добавлять стабилизирующие последовательности, в виде нескольких дополнительных предъявлений в начале первого сеанса, которые не учитываются в результатах оценки (хотя для испытуемого они уже обозначаются как тестовые, а не как тренировочные).

Повторения. Для всех методов, кроме метода PC, рекомендуется вводить как минимум два, а лучше три или четыре повторения идентичных условий (т.е. одного и того же стимула/набора стимулов). В методе PC повторения обычно не используются, поскольку сам метод подразумевает многократное предъявление одного и того же стимула, хотя и в разных парах.

Повторение необходимо включать по ряду причин, но главным образом из-за индивидуальной вариабельности оценок у каждого испытуемого, которую можно оценить и учесть благодаря повторным измерениям. По этому параметру можно, в частности, исключать из выборки ненадежных испытуемых или корректировать процедуру тестирования. Кроме того, повторные измерения помогают компенсировать эффекты обучения и уставания.

Планирование. При планировании экспериментов всегда важно учитывать возможность влияния порядка предъявляемых стимулов. В случае субъективной оценки качества к концу тестовой последовательности испытуемые могут уставать (и давать худшие оценки) или, наоборот, врабатываться и лучше понимать задачу (и давать более адекватные оценки). Для планирования порядка предъявляемых стимулов можно использовать общепринятые стратегии, такие как полная рандомизация, дизайн по латинскому квадрату, дизайн по греко-латинскому квадрату, дизайн по квадрату Юдена, повторяющиеся блоки и другие (Хикс, 1967; Bailey, 2008; Oehlert, 2010; Cunningham, Wallraven, 2011).

Условия просмотра. При публикации результатов рекомендуется указывать все характеристики дисплея, тип используемой видеокарты и прочее. При демонстрации стимулов не во весь экран, а в ограниченном окне экрана, фон должен соответствовать 50%-ному серому. Информацию о том, как определить соответствующие уровни RGB-излучений для конкретного рабочего дисплея, можно найти в работах (Домасёв, Гнатюк, 2007; Wyszecki, Stiles, 2000). Подробнее о параметрах демонстрирующей системы можно посмотреть в рекомендациях (ITU-R P.910, 1999; ITU-T BT.500-13, 2002; ITU-R BT.710, 1998).

Испытуемые (наблюдатели). Согласно рекомендациям (ITU-R P.910, 1999), число участников предварительных серий экспериментов может быть 4–8; в тестовом просмотре обычно участвуют 15–40 испытуемых.

Перед началом экспериментов рекомендуется проверять цветовосприятие (по полихроматическим таблицам) и остроту зрения испытуемых. Указывается, что острота зрения испытуемых должна быть не ниже уровня условной медицин-

ской нормы (с использованием оптической коррекции, если это необходимо), что соответствует 1.0 дес.ед (20/20 по Снеллену).

В описании проводимых экспериментов рекомендуется указывать экспертный уровень участников: необученные наблюдатели, не имеющие опыта оценки искажений изображения, или опытные наблюдатели-эксперты. Важно, что испытуемые не должны быть непосредственными участниками разработки испытываемой системы/способа обработки изображений (ITU-T BT.500-13, 2002).

В полном описании группы рекомендуется также приводить данные о возрастном и половом составе, образовании и профессиональной категории испытуемых (студент, служащий и др.).

Проверка надежности испытуемых. В рекомендациях (ITU-T BT.500-13, 2002; Mantiuk et al., 2012) описана процедура отбора полученных данных для исключения ненадежных измерений: для каждого испытуемого подсчитывается число оценок, которые лежат за пределами двух стандартных отклонений от среднего по группе, и исключаются те испытуемые, у которых доля таких оценок превышает 5% или вероятность получения оценок, отличающихся от среднего более, чем на два стандартных отклонения, не стремится к нулю с увеличением отличия от среднего.

В методах, где испытуемому могут предъявляться рядом стимул и эталон (например, SDSCE, DCR, PCJ), при предъявлении эталонного видео/изображения вместо стимульного испытуемые должны давать оценки, соответствующие полному или почти полному отсутствию различий между стимулами (близкие к 100 в случае SDSCE, близкие к 0 в случае PCJ, близкие к 5 в случае DCR). Это подтверждает, что испытуемые понимают свою задачу и не дают случайных ответов.

Для упрощения сопоставления оценок применяется линейное преобразование, приводящее индивидуальные средние значения и дисперсии к общему среднему значению и дисперсии для всех пользователей (z-преобразование).

Выбор шкалы оценки. Для методов ACR, ACR-HR и DCR по умолчанию рекомендуется использовать ликертовскую[8] 5-балльную шкалу; каждый уровень шкалы имеет свое наименование (“отлично”, “хорошо”, “удовлетворительно, “неудовлетворительно”, “плохо”). Испытуемые могут давать ответы как в вербальной форме, так и в числовых значениях (в этом случае можно использовать как целые числа, так и дроби, но при этом обязательно, чтобы инструкция была единой для всех испытуемых).

Другими рекомендуемыми шкалами являются 9-балльная, 11-балльная и квазинепрерывная шкалы (рис. 9). В квазинепрерывной шкале вербальные обозначения указываются только для конечных значений, в центре шкалы ставится метка, сама шкала делится на некоторое количество дискрет (число дискрет не определено рекомендациями МСЭ (ITU)). Считается, что использование такой шкалы снижает смещения, обусловленные неоднозначностью наименований делений в балльных шкалах. В (ITU-R BT.1082, 1990) представлено сравнение разных шкал и обсуждается преимущество 11-балльной шкалы.

Для социологических и психологических исследований типично рассмотрение шкал ликертовского типа как интервальных. Однако такие шкалы оценок лишь условно интервальные, а по факту являются порядковыми, в связи с чем кажется более предпочтительным указывать в результатах не среднее, а медиану и межквартильный размах; к тому же медиана более устойчива к выбросам в данных, что существенно при малом количестве данных или при несоответствии их распределения нормальному; однако в рассматриваемых документах такой вариант анализа почему-то не обсуждается, а чаще рекомендуется использовать среднее значение.

Окончательный выбор шкалы и разработка инструкции по ее использованию остаются на усмотрение экспериментаторов и должны соответствовать проводимому измерению. Дополнительную информацию можно найти в Приложении В в рекомендациях МСЭ (ITU-R P.910, 1999).

Важно понимать, что в переводе с английского на язык, удобный для испытуемых, словесные наименования одних и тех же категорий шкалы могут немного отличаться по смыслу и иметь разные оттенки, что иногда вносит искажения в результаты, несмотря на стандартизацию процедуры.

Также важно учитывать локальные привычки и практику употребления определенных балльных систем. Например, в России в образовательных учреждениях широко используется пятибалльная шкала оценок. Однако фактически эта шкала редуцирована до 4-балльной: оценка “1” ставится крайне редко, и худшей оценкой является “2” – неудовлетворительно. Использование 5-балльной шкалы при проведении измерений на испытуемых в России представляется нежелательным, так как при этом нередко обрезается нижняя часть диапазона оценок, что приводит к некорректным результатам.

Дополнительные шкалы. К шкале оценки общего качества могут добавляться дополнительные.

[8] Шкала ликертовского типа – балльная (рейтинговая) шкала, позволяющая оценивать мнения испытуемых по нескольким уровням: обычно включает два противоположных мнения на концах шкалы (“отлично” – “плохо”) и имеет нейтральную точку (“удовлетворительно”) в середине; как правило, дополняется вербальными “якорями” вдоль шкалы.

Рис. 9. Шкалы, используемые для оценки впечатлений испытуемых (ITU-R P.910, 1999)
а – ликертовские балльные, б – квазинепрерывная.

В частности, в (ITU-R P.910, 1999) для отдельных показателей качества рекомендуют использовать шесть положительных и пять отрицательных шкал. Положительные шкалы вводят для оценки яркости, контрастности, цветовоспроизведения, очертаний контура, стабильности фона, скорости сборки предъявляемого изображения. По отрицательным шкалам оценивают подергивание, эффекты размытия, эффект москитного шума, двойные изображения/тени, гало. Для оценки изображения по этим показателям рекомендуют использовать квазинепрерывную шкалу. Полученные данные можно объединить с определенными весовыми коэффициентами в общую оценку качества.

Длительности предъявления стимулов. Важными параметрами при планировании эксперимента являются длительность предъявления контента и время голосования. Для видеостимулов рекомендуется ограничивать время предъявления десятью секундами (ITU-R P.910, 1999) (хотя некоторые авторы считают, что достаточно пяти секунд (Mantiuk et al., 2012)), а в случае статических изображений нередко достаточно и трех секунд предъявления (Mantiuk et al., 2012). Если стимулы показываются один за другим, между ними в течение полутора-двух секунд предъявляется пустой серый экран (Mantiuk et al., 2012; ITU-R P.910, 1999). Время для голосования рекомендуется ограничить десятью секундами (ITU-R P.910, 1999).

В документе (ITU-T BT.500-13, 2002) при рассмотрении метода DSCQS обсуждается, что слишком короткие видеопоследовательности можно использовать внутри одного 10-секундного интервала несколько раз. Эти фрагменты можно пристыковывать друг к другу, используя инвертирование видео по времени (для сглаживания стыков). Такой метод демонстрации называют “палиндромным” показом (“palindromic” display) (ITU-T BT.500-13, 2002).

Экспериментальные серии не следует делать слишком длительными, чтобы не утомлять испытуемых, а в сами сессии иногда следует вводить короткие перерывы. В рекомендации (ITU-T BT.500-13, 2002) указано, что продолжительность сеанса не должна превышать получаса (включая объяснения и предварительные этапы).

При планировании экспериментов с видеоконтентом также необходимо учитывать различные психологические эффекты, связанные с временны́ми характеристиками восприятия. В качестве примера можно привести эффект “недавности” (или последних секунд) – recency effect (Kratochvíl,

Slanina, 2010), или эффект “прощения” – forgiveness effect (Alpert, Evian, 1977), заключающиеся в том, что впечатление о хорошем/плохом качестве в первые секунды перебивается более свежим впечатлением о плохом/хорошем качестве в последние секунды. Такие эффекты являются еще одной причиной того, что при тестировании с вынесением однократных оценок предпочтительно использовать короткие видеопоследовательности.

ЗАКЛЮЧЕНИЕ

В настоящем кратком обзоре дано лишь самое общее представление о методологических аспектах исследований, посвященных субъективным методам оценки качества изображений, получаемых в результате различных способов их обработки для повышения эффективности передачи по телекоммуникационным каналам и в других целях. Хотя такие исследования проводятся уже с конца прошлого века, до недавнего времени они не были остро актуальными, поскольку потребность в них была невелика. В связи с этим в отечественной литературе практически нет подробных и полных источников по этой теме. Долгое время проблемы оптимального кодирования, сжатия, фильтрации, трансформации изображений разрабатывались преимущественно в сфере телевидения. При этом большая часть задач по оптимизации процессов передачи и приема визуальной информации решалась с применением объективных математических методов анализа информационных процессов.

За последнее время технический прогресс привел к значительному расширению сферы применения специальных методов оценки качества изображений как в традиционных областях – в системах хранения, передачи и воспроизведения визуальной информации, так и в развивающихся областях визуализации незрительной информации и сравнительной оценки степени полезности эмпирических и полуэмпирических алгоритмов обработки, применяемых в тех сложных случаях, когда аналитически строгий подход невозможен.

Потребность в исследованиях, включающих субъективные оценки качества изображений, в последние годы непрерывно растет, но они будут способствовать общему развитию информационных технологий только при условии корректного проведения экспериментов и адекватного анализа полученных данных. При этом очевидно, что большинство такого рода исследований должны проводиться по стандартным схемам и протоколам, позволяющим сравнивать и интегрировать результаты различных авторов. На данном этапе совершенствование методологии в этой области возможно только на основе накопления достаточно большого количества данных и их сопоставления с результатами применения объективных методов.

Однако логично предположить, что в не столь отдаленном будущем произойдет принципиальное изменение ситуации, и работы по субъективной оценке качества изображений в их классическом виде станут менее востребованными. Во-первых, благодаря получению все более полной информации о механизмах зрительного восприятия человека и созданию достаточно адекватных моделей функционирования зрительной системы при восприятии различных изображений, на основе таких моделей будут разработаны более совершенные объективные методы оценки качества изображений, которые смогут заменить субъективные методы в большинстве случаев. Во-вторых, будет разрабатываться и внедряться все больше систем искусственного зрения, предназначенных как для автономного функционирования, так и в комплексах сенсорного оснащения роботов, в связи с чем алгоритмы обработки изображений и оценки их качества должны будут ориентироваться на “восприятие” технических систем и роботов, а не человека. Системы искусственного зрения могут радикально отличаться от зрительной системы человека, о чем можно судить по уже созданным и успешно работающим техническим вариантам, и о чем косвенно свидетельствует разнообразие вариантов зрительных систем в живой природе. Переориентация на системы искусственного зрения может полностью изменить методологию оценки качества изображений.

СПИСОК ЛИТЕРАТУРЫ

Божкова В.П., Басова О.А., Николаев Д.П. Математические модели пространственного цветовосприятия. *Информационные процессы*. 2019. Т. 19. № 2. С. 187–199.

Боков А.А., Ватолин Д.С. Методика объективной оценки качества восстановления фона в видео. *Цифровая обработка сигналов*. 2016. № 3. С. 26–33.

Ватолин Д.С., Паршин А.Е. Методы для объективной оценки качества видеокодеков по сжатым ими видеопоследовательностям. *Новые информационные технологии в автоматизированных системах*. 2006 № 9. С. 4–12.

Домасёв М.В., Гнатюк С. *Цвет, управление цветом, цветовые расчеты и измерения*. СПб.: “Питер”, 2007. 224 с.

Лебедев Д.С. Модель механизма распознавания ориентации 3-полосных двухградационных оптотипов. *Сенсорные системы*. 2015. Т. 29 № 4. С. 309–320.

Максимов В.В. *Трансформация цвета при изменении освещения*. М.: Наука, 1984. 161 с.

Монич Ю.И., Старовойтов В.В. Оценки качества для анализа цифровых изображений. *Искусственный интеллект*. 2008 № 4. С. 376–385.

Рожкова Г.И., Матвеев С.Г. *Зрение детей: проблемы оценки и функциональной коррекции*. М.: Наука, 2007. 315 с.

Хикс Ч. *Основные принципы планирования эксперимента*. М.: МИР, 1967. 406 с.

Aldridge R.P., Hands D.S., Pearson D.E., Lodge N.K. Continuous quality assessment of digitally coded television pictures. *IEE Proceedings of the Vision Image and Signal Processing*. 1998. V. 145 (2). P. 116–123. https://doi.org/10.1049/ip-vis:19981843

Alpert T., Evain J.-P. Subjective quality evaluation – The SSCQE and DSCQE methodologies. *Ebu Tech Rev*. 1997. P. 12–20.

Anwar S., Li C., Porikli F. Deep underwater image enhancement. *arXiv preprint arXiv:1807.03528*. 2018. 12 p.

Bailey R.A. *Design of Comparative Experiments*. Cambridge University Press. 2008. 255 p. https://doi.org/10.1017/CBO9780511611483

Chen Q., Xu J., Koltun V. Fast image processing with fully-convolutional networks. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. 2017. V. 9. P. 2497–2506.

Corriveau P., Gojmerac C., Hughes B., Stelmach L. All subjective scales are not created equal: The effects of context on different scales. *Signal Processing*. 1999. V. 77 (1). P. 1–9. DOI: (99)00018-3 https://doi.org/10.1016/S0165-1684

Cunningham D.W., Wallraven C. *Experimental design: from user studies to psychophysics*. AK Peters/CRC Press. 2011. 392 p.

van Dijk A.M., Martens J.-B., Watson A.B. Quality assessment of coded images using numerical category scaling. *Proc SPIE*. 1995. V. 2451. P. 90–101. https://doi.org/10.1117/12.201231

Fairchild M.D., Johnson G.M. The iCAM framework for image appearance, image differences, and image quality. *J. Electron. Imaging*. 2004. V. 13. P. 126–138.

Gulliksen H., Tucker L. A general procedure for obtaining paired comparisons from multiple rank orders. *Psychometrika*. 1961. V. 6. P. 173–184.

Hamberg R., de Ridder H. Continuous assessment of perceptual image quality. *Journal of the Optical Society of America*. 1995. V. 12. P. 2573–2577. https://doi.org/10.1364/JOSAA.12.002573

Hamberg R., Ridder H. Time-varying Image Quality: Modeling the Relation between Instantaneous and Overall Quality. *SMPTE Journal*. 1999. P. 802–811. https://doi.org/10.5594/J04337

Hands D.S. A basic multimedia quality model. *IEEE Transactions on Multimedia*. 2004. V. 6 (6). P. 806–816. https://doi.org/10.1109/TMM.2004.837233

Ijsselsteijn W., de Ridder H., Hamberg R., Bouwhuis D., Freeman J. Perceived depth and the feeling of presence in 3DTV. *Displays*. 1998. V. 18. P. 207–214. DOI: (98)00022-5 https://doi.org/10.1016/S0141-9382

ITU-R Recommendation ITU-R P.910. 1999. *Subjective video quality assessment methods for multimedia applications*.

ITU-R Recommendation ITU-T BT.500-13. 2002. *Methodology for the subjective assessment of the quality of television pictures*.

ITU-R Recommendation ITU-R BT.710-4. 1998. *Subjective assessment methods for image quality in high-definition television*.

ITU-R Report ITU-R BT.1082. 1990. *Studies toward the unification of picture assessment methodology*.

ITU-T Recommendation ITU-T P.800. 1996. *Methods for subjective determination of transmission quality*.

ITU-T Recommendation ITU-T P.930. 1996. *Principles of a reference impairment system for video*.

Kim A., Reynaud R., Hess K., Mullen. A normative data set for the clinical assessment of achromatic and chromatic contrast sensitivity using a qCSF approach. *Investigative ophthalmology and visual science*. 2017. V. 58 (9). P. 3628–3636. https://doi.org/10.1167/iovs.17-21645

Kontsevich L.L., Tyler C.W. Analysis of stereothresholds for stimuli below 2.5 c/deg. *Vision Res*. 1994. V. 34 (17). P. 2317–2329. DOI: (94)90110-4 https://doi.org/10.1016/0042-6989

Kontsevich L.L., Tyler C.W. A simpler structure for local spatial channels revealed by sustained perifoveal stimuli. *J. Vis*. 2013. V. 13 (1). P. 1–12. https://doi.org/10.1167/13.1.22

Kratochvíl T., Slanina M. *Digital Video Image Quality, Digital Video*. Floriano De Rango (Ed.). 2010. P. 487–500.

Kuang J., Johnson G.M., Fairchild M.D. iCAM06: A refined image appearance model for HDR image rendering. *Journal of Visual Communication and Image Representation*. 2007. V. 18 (5). P. 406–414. https://doi.org/10.1016/j.jvcir.2007.06.003

Lambooij M., Ijsselsteijn W.A., Heynderickx I. Visual discomfort of 3DTV: Assessment methods and modeling. *Displays*. 2011. V. 32 (4). P. 209–218. https://doi.org/10.1016/j.displa.2011.05.012

Ledig C., Theis L., Huszar F., Caballero J., Cunningham A., Acosta A., Acosta A., Aitken A., Tejani A., Totz J., Wang Z., Shi W. Photo-realistic single image super-resolution using a generative adversarial network. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. 2017. P. 4681–4690.

Li Y., Zhang Y., Xu X., He L., Serikawa S., Kim H. Dust removal from high turbid underwater images using convolutional neural networks. *Opt. Laser Technol*. 2019. V. 110. P. 2–6. https://doi.org/10.1016/j.optlastec.2017.09.017

Lissner I., Preiss J., Urban P., Lichtenauer M.S., Zolliker P. Image-difference prediction: From grayscale to color. *IEEE Transactions on Image Processing*. 2013. V. 22 (2). P. 435–446. https://doi.org/10.1109/TIP.2012.2216279

Lu H., Li Y., Zhang L., Serikawa S. Contrast enhancement for images in turbid water. *J. Optical Society of America A*. 2015. V. 32. P. 886–893. https://doi.org/10.1364/JOSAA.32.000886

Macdiarmid F., Darby P.J. Double-stimulus assessment of television picture quality. *EBU Tech. Rev*. 1982. V. 192. P. 70–79.

Mangeruga M., Bruno F., Cozza M., Agrafiotis P., Skarlatos D. Guidelines for underwater image enhancement based on benchmarking of different methods. *Remote Sensing*. 2018b. V. 10 (10). P. 1652. https://doi.org/10.3390/rs10101652

Mangeruga M., Cozza M., Bruno F. Evaluation of underwater image enhancement algorithms under different environmental conditions. *J. Mar. Sci. Eng*. 2018a. V. 6 (1). P. 1–13. https://doi.org/10.3390/jmse6010010

Mantiuk R.K., Tomaszewska A., Mantiuk R. Comparison of four subjective methods for image quality assessment. *Comput Graph Forum*. 2012. V. 31 (8). P. 2478–2491.
https://doi.org/10.1111/j.1467-8659.2012.03188.x

Mohammadi P., Ebrahimi-Moghadam A., Shirani S. Subjective and Objective Quality Assessment of Image: A Survey. *arXiv preprint arXiv:1406.7799*. 2014. 50 p. http://arxiv.org/abs/1406.7799.

Narita N. Subjective-evaluation method for quality of coded images. *IEEE Trans. Broadcasting*. 1994. V. 40. P. 7–13. https://doi.org/10.1109/11.272416

Narita N., Sugiura Y. On an absolute evaluation method of the quality of television sequences. *IEEE Trans. Broadcasting*. 1997. V. 43. P. 26–35.
https://doi.org/10.1109/11.566821

Oehlert G.W. *A first course in design and analysis of experiments*. University of Minnesota, Minnesota, United States of America. 2010. 659 p.

Pinson M.H., Janowski L., Papir Z. Video Quality Assessment: Subjective testing of entertainment scenes. *IEEE Signal Processing Magazine*. 2015. V. 32 (1). P. 101–114. https://doi.org/10.1109/MSP.2013.2292535

Pinson M.H., Wolf S. Comparing subjective video quality testing methodologies. *Proc. SPIE Visual Communications and Image Processing*. 2003. V. 5150. P. 573–582. https://doi.org/10.1117/12.509908

Phillips J.B., Eliasson H. Subjective Image Quality Assessment–Theory and Practice. *Camera Image Quality Benchmarking*. John Wiley & Sons. 2018. P. 117–166.

Redi J., Liu H., Alers H., Zunino R., Heynderickx I. Comparing subjective image quality measurement methods for the creation of public databases. *Proc. SPIE. Image Quality and System Performance VII*. 2010. V. 7529. P. 752903.
https://doi.org/10.1117/12.839195

Silverstein D., Farrell J. Efficient method for paired comparison. *Journal of Electronic Imaging*. 2001. V. 10. P. 394.

Stoica A., Vertan C., Fernandez-Maloigne C. Objective and subjective color image quality evaluation for JPEG 2000 compressed images. Proc. of International Symposium on Signals, Circuits and Systems. 2003. V. 1. P. 137–140.
https://doi.org/10.1109/SCS.2003.1226967.

Torgerson W.S. *Theory and methods of scaling*. Wiley, 1985. 460 p.

Wang Z., Bovik A.C., Sheikh H.R., Simoncelli E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*. 2004. V. 13 (4). P. 600–612.

Watson A.B., Ahumada A.J. A standard model for foveal detection of spatial contrast. *Journal of Vision*. 2005. V. 5 (9). P. 717–740.
https://doi.org/10.1167/5.9.6

Watson A.B., Ahumada A.J. Predicting visual acuity from wavefront aberrations. *Journal of Vision*. 2008. V. 8 (4). P. 1–19.
https://doi.org/10.1167/8.4.17

Watson A.B., Ahumada A.J. Modeling acuity for optotypes varying in complexity. *Journal of Vision*. 2012. V. 12 (10). P. 1–19.
https://doi.org/10.1167/12.10

Watson A., Ramirez C. V., Salud E. Predicting visibility of aircraft. *PloS one*. 2009. V. 4 (5). C. e5594.
https://doi.org/10.1371/journal.pone.0005594

Wyszecki G., Stiles W.S. *Color science. Concepts and Methods. Quantitative data and Formulae*. Wiley. 2000. 950 p.

Xie Z.-X., Wang Z.-F. Color image quality assessment based on image quality parameters perceived by human vision system. *Proc. of the 2010 IEEE International Conference on Multimedia Technology* (*ICMT*). 2010. P. 1–4.
https://doi.org/10.1109/ICMULT.2010.5630949

Xu L., Lin W., Kuo C.C.J. *Visual quality assessment by machine learning*. Springer Singapore. 2015. 132 p.
https://doi.org/10.1007/978-981-287-468-9

Yano S., Ide S., Mitsuhashi T., Thwaites H. A study of visual fatigue and visual comfort for 3D HDTV/HDTV images. *Displays*. 2002. V. 23. P. 191–201.
DOI: (02)00038-0
https://doi.org/10.1016/S0141-9382

Zhang X., Wandell B.A. A spatial extension of CIELAB for digital color image reproduction. *SID International Symposium Digest of Technical Papers*: *Society for Information Display*. 1996. V. 27.P. 731–734.
https://doi.org/10.1889/1.1985127

Zhang X., Wandell B.A. Color image fidelity metrics evaluated using image distortion maps. *Signal Processing*. 1998. V. 70 (3). P. 201–214. DOI: (98)00125-X
https://doi.org/10.1016/S0165-1684

Subjective image and video quality assessment: methodology review

M. A. Gracheva[a,#], V. P. Bozhkova[a], A. A. Kazakova[a,b], and G. I. Rozhkova[a]

[a] *Institute for Information Transmission Problems (Kharkevich Institute), Russian Academy of Sciences, 127051 Moscow, B. Karetny per. 19, Build. 1, Russia*

[b] *Pirogov Russian National Research Medical University, 117997 Moscow, Ostrovitianov str. 1, Russia*

[#] *E-mail: mg.iitp@gmail.com*

Usability of different algorithms elaborated for visual information compression, filtering or transformation is usually compared by means of various methods of assessing the quality of the images obtained as a result of processing. There are two types of such methods – objective and subjective. Objective methods are based on strict mathematical criteria independent of people's opinions. In contrast, subjective methods are based on the people's opinions collected in special experiments on image quality assessment. In this brief review, some methodological aspects of the subjective methods are considered. Though the assessment of image quality by means of such methods is in high demand at present, there is little proper literature in Russian for the potential users, having no experience in such investigations. The methods succinctly described here are: ACR – absolute category rating, ACR-HR – absolute category rating with hidden reference, SSCQE – single stimulus

continuous quality estimation, DCR – degradation category rating, DSCQR – double stimulus continuous quality rating, PC – pair comparison, PSJ – pairwise similarity judgement, SDSCE – simultaneous double stimulus for continuous evaluation. In addition, some general recommendations are presented on planning and conducting of the experiments, implying participation of many people.

Key words: image quality, subjective image quality assessment, human studies, user studies, experimental design

REFERENCES

Bozhkova V.P., Basova O.A., Nikolaev D.P. Matematicheskie modeli prostranstvennogo tsvetovospriyatiya [Mathematical models of spatial color perception]. *Informatsionnye protsessy* [Information processes]. 2019. V. 19. № 2. P. 187–199 (in Russian).

Bokov A.A., Vatolin D.S. Metodika ob"ektivnoi otsenki kachestva vosstanovleniya fona v video [Objective quality assessment methodology for video background reconstruction]. *Tsifrovaya obrabotka signalov* [Digital Signal Processing]. 2016 № 3. P. 26–33 (in Russian).

Vatolin D.S., Parshin A.E. Metody dlya ob"ektivnoi otsenki kachestva videokodekov po szhatym imi videoposledovatel'nostyam [Methods for an objective assessment of the quality of video codecs using video sequences compressed by them]. *Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh* [New information technologies in automated systems]. 2006 № 9. P. 4–12 (in Russian).

Domasev M.V., Gnatyuk S. *Tsvet, upravlenie tsvetom, tsvetovye raschety i izmereniya* [Color, color management, color calculations and measurements]. SPb.: "Piter", 2007. 224 p. (in Russian).

Lebedev D.S. Model' mekhanizma raspoznavaniya orientatsii 3-polosnykh dvukhgradatsionnykh optotipov [A model of orientation recognition mechanisms for the 3-bar two-grade optotypes]. *Sensornye sistemy* [Sensory systems]. 2015. V. 29 (4). P. 309–320 (in Russian).

Maksimov V.V. *Transformatsiya tsveta pri izmenenii osveshcheniya* [Color transformation during lighting changes]. M.: Nauka, 1984. 161 p. (in Russian).

Monich Yu.I., Starovoitov V.V. Otsenki kachestva dlya analiza tsifrovykh izobrazhenii [Image Quality Evaluation for Image Analysis]. *Iskusstvennyi intellekt* [Artificial intelligence]. 2008. (4). P. 376–385 (in Russian).

Rozhkova G.I., Matveev S.G. *Zrenie detei: problemy otsenki i funktsional'noi korrektsii* [Children vision: problems of vision assessment and functional treatment]. M.: Nauka, 2007. 315 p. (in Russian).

Khiks Ch. *Osnovnye printsipy planirovaniya eksperimenta* [Basic principles of experiment planning]. M.: MIR, 1967. 406 p. (in Russian).

Aldridge R.P., Hands D.S., Pearson D.E., Lodge N.K. Continuous quality assessment of digitally coded television pictures. *IEE Proceedings of the Vision Image and Signal Processing*. 1998. V. 145 (2). P. 116–123. DOI: 10.1049/ip-vis:19981843

Alpert T., Evain J-P. Subjective quality evaluation – The SSCQE and DSCQE methodologies. *Ebu Tech Rev*. 1997. P. 12–20.

Anwar S., Li C., Porikli F. Deep underwater image enhancement. *arXiv preprint arXiv:1807.03528*. 2018. 12 p.

Bailey R.A. *Design of Comparative Experiments*. Cambridge University Press. 2008. 255 p. DOI: 10.1017/CBO9780511611483

Chen Q., Xu J., Koltun V. Fast image processing with fully-convolutional networks. Proc. *IEEE Conference on Computer Vision and Pattern Recognition*. 2017. V. 9. P. 2497–2506.

Corriveau P., Gojmerac C., Hughes B., Stelmach L. All subjective scales are not created equal: The effects of context on different scales. *Signal Processing*. 1999. V. 77 (1). P. 1–9. DOI: 10.1016/S0165-1684 (99)00018-3

Cunningham D.W., Wallraven C. *Experimental design: from user studies to psychophysics*. AK Peters/CRC Press. 2011. 392 p.

van Dijk A.M., Martens J.-B., Watson A.B. Quality assessment of coded images using numerical category scaling. *Proc SPIE*. 1995. V. 2451. P. 90–101. DOI: 10.1117/12.201231

Fairchild M.D., Johnson G.M. The iCAM framework for image appearance, image differences, and image quality. *J. Electron. Imaging*. 2004. V. 13. P. 126–138.

Gulliksen H., Tucker L. A general procedure for obtaining paired comparisons from multiple rank orders. *Psychometrika*. 1961. V. 26. P. 173–184.

Hamberg R., de Ridder H. Continuous assessment of perceptual image quality. *Journal of the Optical Society of America*. 1995. V. 12. P. 2573–2577. DOI: 10.1364/JOSAA.12.002573

Hamberg R., Ridder H. Time-varying Image Quality: Modeling the Relation between Instantaneous and Overall Quality. *SMPTE Journal*. 1999. P. 802–811. DOI: 10.5594/J04337

Hands D.S. A basic multimedia quality model. *IEEE Transactions on Multimedia*. 2004. V. 6 (6). P. 806–816. DOI: 10.1109/TMM.2004.837233

Ijsselsteijn W., de Ridder H., Hamberg R., Bouwhuis D., Freeman J. Perceived depth and the feeling of presence in 3DTV. *Displays*. 1998. V. 18. P. 207–214. DOI: 10.1016/S0141-9382 (98)00022-5

ITU-R Recommendation ITU-R P.910. 1999. *Subjective video quality assessment methods for multimedia applications.*

ITU-R Recommendation ITU-T BT.500-13. 2002. *Methodology for the subjective assessment of the quality of television pictures.*

ITU-R Recommendation ITU-R BT.710-4. 1998. *Subjective assessment methods for image quality in high-definition television.*

ITU-R Report ITU-R BT.1082. 1990. *Studies toward the unification of picture assessment methodology.*

ITU-T Recommendation ITU-T P.800. 1996. *Methods for subjective determination of transmission quality.*

ITU-T Recommendation ITU-T P.930. 1996. *Principles of a reference impairment system for video.*

Kim A., Reynaud R., Hess K., Mullen. A normative data set for the clinical assessment of achromatic and chromatic contrast sensitivity using a qCSF approach. *Investigative ophthalmology and visual science*. 2017. V. 58 (9). P. 3628–3636. DOI:10.1167/iovs.17-21645

Kontsevich L.L., Tyler C.W. Analysis of stereothresholds for stimuli below 2.5 c/deg. *Vision Res*. 1994. V. 34 (17). P. 2317–2329. DOI: 10.1016/0042-6989 (94)90110-4

Kontsevich L.L., Tyler C.W. A simpler structure for local spatial channels revealed by sustained perifoveal stimuli. *J. Vis*. 2013. V. 13 (1). P. 1–12. DOI: 10.1167/13.1.22

Kratochvíl T., Slanina M. *Digital Video Image Quality, Digital Video*. Floriano De Rango (Ed.). 2010. P. 487–500.

Kuang J., Johnson G.M., Fairchild M.D. iCAM06: A refined image appearance model for HDR image rendering. *Journal of Visual Communication and Image Representation*. 2007. V. 18 (5). P. 406–414. DOI: 10.1016/j.jvcir.2007.06.003

Lambooij M., Ijsselsteijn W.A., Heynderickx I. Visual discomfort of 3DTV: Assessment methods and modeling. *Displays*. 2011. V. 32 (4). P. 209–218. DOI: 10.1016/j.displa.2011.05.012

Ledig C., Theis L., Huszar F., Caballero J., Cunningham A., Acosta A., Acosta A., Aitken A., Tejani A., Totz J., Wang Z., Shi W. Photo-realistic single image super-resolution using a generative adversarial network. Proc. *IEEE Conference on Computer Vision and Pattern Recognition*. 2017. P. 4681–4690.

Li Y., Zhang Y., Xu X., He L., Serikawa S., Kim H. Dust removal from high turbid underwater images using convolutional neural networks. *Opt. Laser Technol*. 2019. V. 110. P. 2–6. DOI: 10.1016/j.optlastec.2017.09.017

Lissner I., Preiss J., Urban P., Lichtenauer M.S., Zolliker P. Image-difference prediction: From grayscale to color. *IEEE Transactions on Image Processing*. 2013. V. 22 (2). P. 435–446. DOI: 10.1109/TIP.2012.2216279

Lu H., Li Y., Zhang L., Serikawa S. Contrast enhancement for images in turbid water. *J. Optical Society of America A*. 2015. V. 32. P. 886–893. DOI: 10.1364/JOSAA.32.000886

Macdiarmid F., Darby P. J. Double-stimulus assessment of television picture quality. EBU Tech. Rev. 1982. V. 192. P. 70–79.

Mangeruga M., Bruno F., Cozza M., Agrafiotis P., Skarlatos D. Guidelines for underwater image enhancement based on benchmarking of different methods. Remote Sensing. 2018b. V. 10 (10). P. 1652. DOI: 10.3390/rs10101652

Mangeruga M., Cozza M., Bruno F. Evaluation of underwater image enhancement algorithms under different environmental conditions. *J. Mar. Sci. Eng*. 2018a. V. 6 (1). P. 1–13. DOI: 10.3390/jmse6010010

Mantiuk R.K., Tomaszewska A., Mantiuk R. Comparison of four subjective methods for image quality assessment. *Comput Graph Forum*. 2012. V. 31 (8). P. 2478–2491. DOI:10.1111/j.1467-8659.2012.03188.x

Mohammadi P., Ebrahimi-Moghadam A., Shirani S. Subjective and Objective Quality Assessment of Image: A Survey. *arXiv preprint arXiv:1406.7799*. 2014. 50 p. http://arxiv.org/abs/1406.7799.

Narita N. Subjective-evaluation method for quality of coded images. *IEEE Trans. Broadcasting*. 1994. V. 40. P. 7–13. DOI: 10.1109/11.272416

Narita N., Sugiura Y. On an absolute evaluation method of the quality of television sequences. *IEEE Trans. Broadcasting*. 1997. V. 43. P. 26–35. DOI: 10.1109/11.566821

Oehlert G.W. A first course in design and analysis of experiments. *University of Minnesota, Minnesota, United States of America*. 2010. 659 p.

Pinson M.H., Janowski L., Papir Z. Video Quality Assessment: Subjective testing of entertainment scenes. *IEEE Signal Processing Magazine*. 2015. V. 32 (1). P. 101–114. DOI: 10.1109/MSP.2013.2292535

Pinson M.H., Wolf S. Comparing subjective video quality testing methodologies. Proc. *SPIE Visual Communications and Image Processing*. 2003. V. 5150. P. 573–582. DOI: 10.1117/12.509908

Phillips J.B., Eliasson H. Subjective Image Quality Assessment–Theory and Practice. *Camera Image Quality Benchmarking*. John Wiley & Sons. 2018. P. 117–166.

Redi J., Liu H., Alers H., Zunino R., Heynderickx I. Comparing subjective image quality measurement methods for the creation of public databases. *Proc. SPIE. Image Quality and System Performance VII*. 2010. V. 7529. P. 752903. DOI: 10.1117/12.839195

Silverstein D., Farrell J. Efficient method for paired comparison. *Journal of Electronic Imaging*. 2001. V. 10. P. 394.

Stoica A., Vertan C., Fernandez-Maloigne C. Objective and subjective color image quality evaluation for JPEG 2000 compressed images. *Proc. of International Symposium on Signals, Circuits and Systems*. 2003. V. 1. P. 137–140. DOI: 10.1109/SCS.2003.1226967

Torgerson W.S. *Theory and methods of scaling*. Wiley, 1985. 460 p.

Wang Z., Bovik A.C., Sheikh H.R., Simoncelli E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*. 2004. V. 13 (4). P. 600–612.

Watson A.B., Ahumada A.J. A standard model for foveal detection of spatial contrast. *Journal of Vision*. 2005. V. 5 (9). P. 717–740. DOI: 10.1167/5.9.6.

Watson A.B., Ahumada A.J. Predicting visual acuity from wavefront aberrations. Journal of Vision. 2008. V. 8(4). P. 1–19. DOI: 10.1167/8.4.17.

Watson A.B., Ahumada A.J. Modeling acuity for optotypes varying in complexity. *Journal of Vision*. 2012. V. 12 (10). P. 1–19. DOI: 10.1167/12.10.19.

Watson A.B., Ramirez C.V., Salud E. Predicting visibility of aircraft. PloS one. 2009. V. 4(5). C. e5594.

Wyszecki G., Stiles W.S. Color science. Concepts and Methods. *Quantitative data and Formulae*. Wiley. 2000. 950 p.

Xie Z.-X., Wang Z.-F. Color image quality assessment based on image quality parameters perceived by human vision system. *Proc. of the 2010 IEEE International Conference on Multimedia Technology* (*ICMT*). 2010. P. 1–4. DOI: 10.1109/ICMULT.2010.5630949

Xu L., Lin W., Kuo C.C.J. *Visual quality assessment by machine learning*. Springer Singapore. 2015. 132 p. DOI: 10.1007/978-981-287-468-9

Yano S., Ide S., Mitsuhashi T., Thwaites H. A study of visual fatigue and visual comfort for 3D HDTV/HDTV images. *Displays*. 2002. V. 23. P. 191–201. DOI: 10.1016/S0141-9382 (02)00038-0

Zhang X., Wandell B.A. A spatial extension of CIELAB for digital color image reproduction. *SID International Symposium Digest of Technical Papers: Society for Information Display*. 1996. V. 27. P. 731–734. DOI: 10.1889/1.1985127

Zhang X., Wandell B.A. Color image fidelity metrics evaluated using image distortion maps. *Signal Processing*. 1998. V. 70 (3). P. 201–214. DOI: 10.1016/S0165-1684 (98)00125-X