УДК 004.932

СТАТИСТИЧЕСКАЯ НЕДОСТОВЕРНОСТЬ РАСПРОСТРАНЁННЫХ КРИТЕРИЕВ ОЦЕНКИ КАЧЕСТВА ИСКАЖЁННОГО ИЗОБРАЖЕНИЯ

Голованов Р.В., *аспирант Национального исследовательского университета «МИЭТ», г. Зеленоград, e-mail: golovanovrv@gmail.com;*
Калиткин Н.Н., *член-корреспондент института прикладной математики им. М.В. Келдыша РАН, г. Москва, e-mail: kalitkin@imamod.ru.*

Ключевые слова: обработка изображений, критерии качества, базы тестовых изображений.

Проведён критический анализ наиболее популярных баз тестовых изображений TID2008 и LIVE. Предложен метод объединения этих баз. Проведено сравнение 19 различных критериев оценки качества искажённых изображений. Показано, что сравнение критериев по коэффициентам корреляции Пирсона, Спирмена и Кендалла является не информативным. Предложено проводить сравнение по отношению стандартных уклонений базы и критерия. Проведено ранжирование критериев по этому правилу и показано, что даже лучшие критерии статистически недостоверны.

Введение

При передаче изображения по каналам связи возникают различные искажения. Часть из них связана с техническими сбоями каналов (в числе которых может быть передача по воздуху). Другая часть намеренно вводится при сжатии изображения, если требуется существенно уменьшить количество передаваемой информации. В обоих случаях в искаженном изображении вместо истинной яркости n-го пикселя x_n будет стоять изменённое число y_n. Пользователь должен судить об исходном изображении по полученному искаженному изображению.

Разработка, как аппаратуры, так и способов сжатия с потерями направлены на получение достаточно хорошего качества переданного изображения. При тестировании такой работы можно предложить пользователю большой набор исходных изображений и переданных искажённых изображений, чтобы пользователь визуально оценил качество. Для аккуратного тестирования необходимо привлекать большую группу испытуемых и чётко регламентировать процедуру сравнения. Это очень трудоёмкая и дорогостоящая работа, аналогичная экспериментам в физике и технике. Для каждого нового устройства или нового метода обработки изображений такую процедуру необходимо повторять заново. Поэтому на практике действуют иначе. Создают базу эталонных и искаженных изображений и один раз выполняют экспертную оценку качества изображений (MOS – mean opinion score). Далее по массивам x_n и y_n пытаются построить некоторый индекс качества – математическую функцию от массивов яркостей $f(X,Y)$. Если ухудшению экспертной оценки соответствует убывание этой функции, то данным индексом качества можно пользоваться при конструкторско-исследовательской работе вместо проведения новых экспертных оценок.

Актуальной проблемой является построение формального критерия оценки качества искажённого изображения. Сейчас известны десятки таких критериев, и этому вопросу посвящена обширная литература [1-10 и др.]. Первоначально исследователи исходят из некоторых естественных идей и получают достаточно простые критерии. На практике они не очень хорошо совпадают с экспертными оценками. Их начинают усложнять, нередко эклектичным образом, и при этом вводят ряд свободных (подгоночных) параметров. Итоговые формулы зачастую оказываются весьма громоздкими, а их адекватность человеческому восприятию – не очевидной. Здесь мы рассмотрим 19 критериев наиболее популярных в научно-исследовательских работах.

Известно несколько баз для тестирования самих критериев [1, 2 и др.]. Эти базы отличаются не только набором исходных искажённых изображений, но и принципом получения экспертных оценок. Поэтому один и тот же критерий на разных базах может получать не одинаковые средневзвешенные оценки, а ранжирование группы критериев оказывается зависящим от базы. Эта проблема недостаточно исследована в литературе. Требуется проведение тщательного анализа самих баз: насколько хорошо структурированы данные, каковы стандарты приведенных чисел (статистический разброс экспертных оценок), насколько чётко формулировались инструкции экспертам и проверялась ли правильность их понимания.

Тестовые базы

Рассмотрим наиболее популярные базы тестовых изображений, которые являются публично доступными.

База TID2008 является самой объёмной по количеству тестовых изображений. Она опубликована в 2008 году и сейчас готовится её новая версия. База содержит 25 оригинальных изображений (24 естественных изображения из базы Кодак [3] и одно искусственное). Для каждого оригинала брались 17 искажений различных типов. Каждое искажение представлено 4-мя уровнями интенсивности, соответствующие примерно 21, 24, 27, 30 дБ по критерию PSNR. Таким образом, общее количество изображений в базе 1700.

Типы представленных искажений многообразны.

Здесь содержатся несколько видов шумовых помех, артефакты от сжатия, дефекты передачи по каналам связи, искажения в результате цветовой коррекции и некоторые специфические типы. Многие типы искажения встречаются только в этой базе. Кроме того разработчики базы хорошо упорядочили весь материал. Это легко позволяет выделять и отдельно обрабатывать изображения, группируя их по оригиналам, типам искажений и даже интенсивностям.

Для получения экспертных оценок использовался метод попарного сравнения искажённых изображений для данного оригинала. Каждому эксперту предлагалось из двух искажённых изображений выбрать то, которое меньше отличается от оригинала. Для получения результата необходимо было выполнить попарное сравнение 68-ми тестовых изображений, то есть проделать 2210 сравнений по каждому оригиналу. Но человеку трудно выполнить такое число сравнений. Поэтому использовалась Швейцарская система оценивания (она часто используется при проведении шахматных турниров). Предварительно 68 искажённых изображений разбили случайным образом на 34 пары и показали их по очереди данному респонденту. Изображению, «выигравшему» в сравнении, зачислялся 1 балл. Далее в парном сравнении участвовали изображения с одинаковым количеством баллов и так 9 раз. Каждый эксперт сделал по 34x9=306 парных сравнений.

В проведение эксперимента было привлечено 838 участников из разных стран. Каждое тестовое изображение в среднем оценили 33 эксперта. При обработке первичных данных были исключены некоторые результаты, сильно выбивавшиеся из общей картины. Окончательные экспертные оценки представлены в шкале от 0 до 9, где 0 соответствует наихудшему качеству, а 9 – наилучшему. В базе приведена дисперсия MOS, равная 0.63 балла. Поскольку средняя оценка получалась по 33 экспертным, стандартное отклонение средней оценки равно 0.140 балла. Эта очень важная информация; не все базы её содержат.

Укажем недостатки базаTID2008, в том числе и отмечаемые её авторами [4].

1. Отсутствуют суперпозиции искажений разных типов, хотя в реальной жизни приходится иметь дело именно с суперпозициями.

2. Эксперты не всегда правильно понимали инструкцию. Особенно это заметно на искажениях контраста. Изображения с 20-40% увеличением контраста выглядели лучше оригинала, и эксперты повышали им оценку вместо занижения.

3. Следовало бы дать не среднюю дисперсию по всей базе, а дисперсии (стандарты) для каждого искажённого изображения отдельно. Это позволило бы применять более достоверные статистические обработки.

База LIVE– одна из первых тестовых баз, предложенная в 2006 году. Основой для базы стали 29 изображений с различным содержанием. Для каждого оригинала в среднем предлагается по 26 искажённых изображений. В таких изображениях присутствует один тип искажения из 5 наиболее часто употребимых (сжатия JPEG и JPEG2000, белый шум, гауссово размытие и смазывание).

В получении экспертных оценок было задействовано 138 человек. Большинство из них имеют образование в области цифровой обработки сигналов. Каждый участник оценивал искажённое изображение в баллах от 1 (очень хорошее) до 100 (очень плохое). При этом на экране одновременно показывалось лишь одно искажённое изображение, а оригинал не предоставлялся. Это психологически затрудняло эксперту выставление количественной оценки, несмотря на предварительный тренинг. Каждый эксперт физически не мог оценить все 779 искажений. Поэтому эксперимент проводился в 7 сессий: по две для JPEG и JPEG2000, по одной для других типов искажений. Каждому искажённому изображению приписывалась усреднённая по всем экспертам оценка.

Предварительно проводилась обычная статистическая обработка. Исключались отдельные экспертные оценки, отличие которых от средних были статистически недостоверным. В очищенном материале на каждое искаженное изображение приходится в среднем по 30 экспертных оценок. Стандартное отклонение составляет 2.33 балла. Эта база по-прежнему пользуется спросом во многих исследовательских работах. Основные причины: наличие самых востребованных типов искажений и сравнительно большой объём базы.

Отметим ряд недостатков базы LIVE.

1. Материал структурирован хуже, чем в базе TID2008. Можно группировать по эталонным изображениям и типам искажений, но нельзя сгруппировать по интенсивностям.

2. Количество типов искажений не велико и отсутствуют их суперпозиции.

3. Не приведены стандарты для каждой усреднённой оценки.

Другие базы пока ещё не получили широкого распространения, но уже встречаются в некоторых исследовательских работах. Здесь коротко опишем основные параметры баз Cornell-A57 [5], IVC [6], Toyama-MICT [7] и CSIQ [8]. Почти все они уступают базам LIVE и TID2008 по количеству рассматриваемых типов искажений и по количеству тестовых изображений. База Cornell-A57 содержит всего 54 тестовых изображения с 6 стандартными видами искажений. В базе IVC предложено 185 тестовых изображений 4-х видов искажений, полученных по 10 эталонам. Toyama-MICT состоит из 168 тестовых изображений, полученных как результат сжатия JPEGи JPEG2000. Основой для базы CSIQ служат 30 эталонных изображений. Для каждого из них были получены 4-5 тестовых изображения для 6 различных видов искажений, что составляет 866 тестовых изображений. Видно, что все четыре базы по числу типов искажений заметно уступают базе TID2008. При этом три из них по объёму заметно уступают базе LIVE. Поэтому далее мы будем работать только на базах LIVE и TID2008.

Объединение баз

Проблема

Критерии качества следует тестировать не на одной базе, а на возможно большем числе баз. Это даёт более объективную оценку критерию. Обычно разработчики

критериев приводят значения коэффициентов корреляции на каждой из баз и по этим результатам делают выводы. Зачастую один и тот же критерий в сравнении с другими оказывается на разных местах в разных базах после ранжирования. Заметное отличие в занимаемых местах (дисперсия номера места) указывает на нестабильность критерия. Зависимость MOS от критерия можно проиллюстрировать на точечном графике. Чем лучше облако точек похоже на узкую монотонную кривую, тем адекватнее и стабильнее работает критерий. Критерий VIF [9] был признан лучшим по результатам наших прошлых работ [10]. Структуры облаков в базе TID2008 и LIVE показывают, что даже для лучшего критерия зависимость не является функцией. Тем не менее, облака MOS(VIF) в обеих базах подобны друг другу.

Попытки окончательного ранжирования одновременно по всем базам раньше предпринимались [10, 11], но оказались не вполне удачными. Итоговое ранжирование удобнее сделать, если предварительно объединить все тестовые базы в одну (совместить облака точек друг с другом). Далее мы опишем методику объединения на примере баз TID2008 и LIVE по критерию VIF.

АнализTID2008

Первый шаг – это анализ облака точек для зависимости MOS(VIF). График этого облака приведён на рис. 1. Точки основного скопления отмечены маркером «×». Скопление похоже на эллипс. Ниже эллипса в средней части графика бросается в глаза небольшое скопление точек, обозначенных маркером «○». Для критерия VIF это единственное исключение. Отметим, что есть ряд критериев, которые имеют несколько скоплений точек, выбивающихся из основной массы.

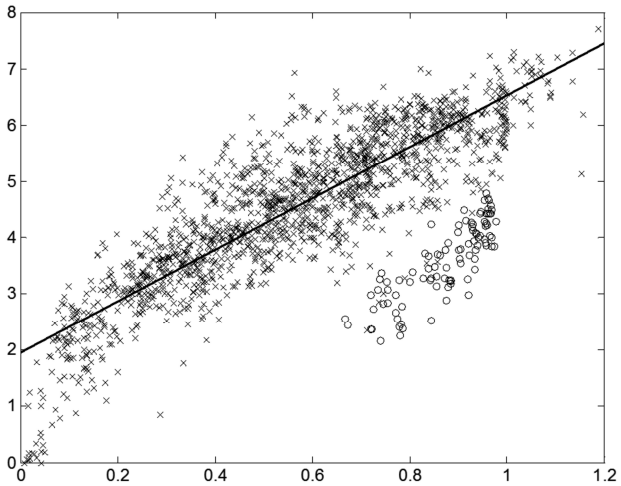

Рис. 1 .Критерий VIF на базе TID2008.
○ – искажения типа №15, × – все остальные типы искажения, прямая – линейная регрессия

По неоднородности облака точек можно судить об адекватности критерия на отдельных типах искажений. Для этого достаточно оценить, как накладываются точки данного искажения на точки всех других искажений. По критерию VIF было выявлено, что обособленное скопление соответствует только изображениям с локальными блочными искажениями разной интенсивности (тип искажения №15 из спецификации базы). Точки других искажений лежат в основном облаке. В данной базе каждому типу искажения соответствует 4 уровня интенсивности. Для ряда критериев по некоторым типам искажения наличие этих уровней можно проследить на графике. Это будет выглядеть, как несколько небольших «пятнышек», лежащих на монотонной кривой. В таком случае критерий можно считать хорошо работающим с данным типом искажения.

Далее оценивалось среднеквадратичное отклонение точек от регрессионной кривой. При этом исключались точки, не вошедшие в основное облако. Среднеквадратичные отклонения для константной, линейной и параболической регрессий оказались равными, соответственно 1.35, 0.67, и 0.64. Видно, что константная регрессия существенно хуже линейной, а отличие линейной регрессии от параболической статистически незначимо. Поэтому линейная регрессия

$$f_{TID\,2008}(x) = 4.5800 \cdot x + 1.9550, \quad (1)$$

является оптимальной. Далее будем использовать только её.

Анализ LIVE

В этой базе типы искажений не так многообразны как TID2008. Поэтому здесь для большинства критериев облако точек выглядит достаточно однородным, без резких выбросов. На рис. 2 приведён график MOS(VIF) для базы LIVE. Видно, что облако точек имеет такую же форму, что и на рис. 1. Напомним, что здесь большим значениям MOS соответствует худшее качество изображения, поэтому облако ориентировано иначе. По этим точкам строились регрессии аналогичным образом. Их среднеквадратичные отклонения соответственно составили 16.1, 5.5 и 5.1.Оптимальной также оказалась регрессия

$$f_{LIVE}(x) = -54.4225 \cdot x + 67.4189\,. \quad (2)$$

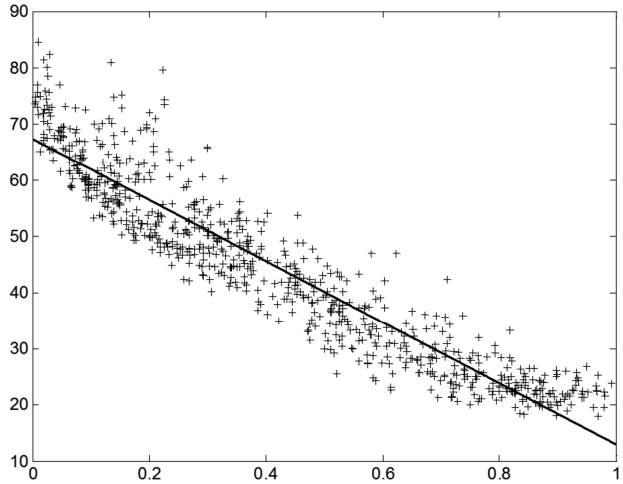

Рис. 2. Критерий VIF на базе LIVE. + – все искажения, прямая – линейная регрессия

Объединение баз сводится к преобразованию шкалы экспертных оценок нескольких баз к одной. Это преобразование должно быть таким, чтобы оптимальные

регрессионные кривые точно совпадали. Для линейных регрессий (1) и (2) искомое преобразование будет линейным:

$$X' = -0.084157 \cdot X + 7.6287, \quad (3)$$

где X – экспертные оценки LIVE в «родной» шкале, а X' – в шкале TID2008.

Результат объединения показан на рис. 3. Видно, что основное облако стало плотнее в сравнении с рис. 1. Визуально совмещение баз сделано хорошо. Формально в этом можно убедиться, вычислив линейную регрессию для нового, более плотного облака. Новая регрессия точно совпадают с (1).

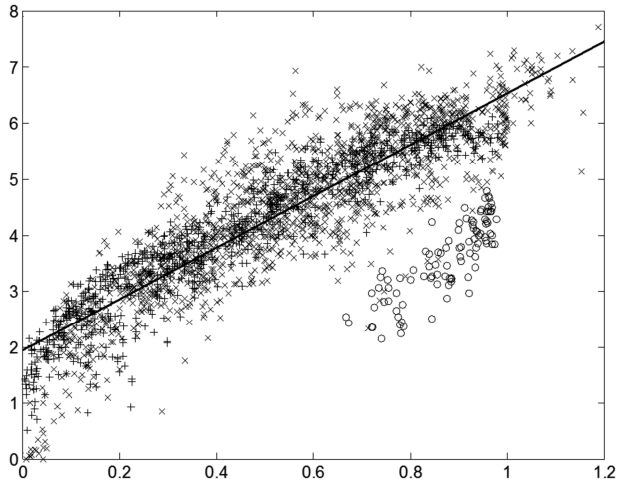

Рис. 3. Критерий VIF на объединенной базе. Маркеры см. рис. 1 и 2

Графики для других критериев на объединённой базе повторили результат из рис. 3. Основные облака хорошо накладываются друг на друга. Однако были выявлены типы искажений, на которых многие критерии работают плохо. Скопления точек для этих искажений не укладывалось в основное облако. Большинство критериев дают не адекватный результат на искажениях №12 (сбои в канале передачи JPEG изображения), 14 (перестановки кусков изображения), 15 (наложение однотонных квадратов), 16 (изменение яркости) и 17 (изменение контрастности) из базы TID2008. С этими искажениями не справляются в среднем более половины рассматриваемых критериев.

Такое поведение обусловлено не стабильным экспертным мнением по данным типам искажений. В зависимости от содержания изображений для данной интенсивности искажения получены сильно отличающиеся значения MOS. Для равноправного сравнения работы метрик качества эти группы следует исключать из рассмотрения.

Были проведены аналогичные (очень трудоёмкие) работы по объединению баз на основе других критериев. Для них картины облаков точек оказываются существенно более сложными (например, кометообразными и содержащими несколько боковых выбросов), а плотность распределения точек в облаках была заметно не равномерной. Для ряда критериев облака на базах TID2008 и LIVE сильно отличались друг от друга. Во всех этих случаях не представлялось разумных регрессий, совмещающих облака обеих баз. Объединение баз с помощью таких критериев не разумно. Именно поэтому мы здесь представляем объединение баз только по критерию VIF.

Важной информацией для любой базы является наличие стандартного отклонения для экспертных оценок. MOS базы TID2008 равен 0.140, а для базы LIVE после преобразования (3) составляет 0.194. Стандарт объединённой базы вычисляется по правилам статистики:

$$S = \sqrt{\frac{S_1^2 \cdot N_1 + S_2^2 \cdot N_2}{N_1 + N_2}},$$

где S_1 и S_2 – стандарт MOS одной и другой баз, а N_1 и N_2 количество тестовых изображений в них соответственно. Для объединённой базыTID2008(со всеми искажениями)+LIVE получаем $S = 0.160$. Исключая из объединённой базы пяти «плохих» групп TID2008, получаем усечённую базу с $S = 0.164$. Оба стандарта будем использовать при анализе результатов.

Тестирование критериев

Будем тестировать критерии на объединённой базе. Для этого обычно используют коэффициенты парной корреляции между значениями критерия и MOS на данной базе. Используют коэффициент линейной корреляции Пирсона PLCC, ранговой Спирмена SRCC и Кендалла KRCC. Чем ближе значение к 1.0, тем адекватнее критерий человеческому восприятию. По этим коэффициентам данный критерий сравнивают с другими. При этом ранжирование критериев по разным коэффициентам корреляции оказывается не одинаковым.

В [10] отмечалось, что наиболее жёсткой оценкой является сравнение по PLCC, а сравнения по SRCC и KRCC менее показательны. Сейчас мы пришли к выводу, что сравнение по всем коэффициентам корреляции не является информативным. В самом деле, достаточно смоделировать работу критерия. В плоскости XY возьмём параллелограмм с границами $x = \pm 1$, $y = x \pm \alpha$. Равномерно заполним его большим количеством точек. Для такого облака оптимальной будет регрессия $y(x) = x$ со стандартным уклонением $\alpha / \sqrt{3}$ и коэффициентом парной корреляции Пирсона $\left(1 + \alpha^2\right)^{-1/2}$. Для $\alpha = 0.3$ и 0.5 коэффициенты корреляции соответственно равны 0.96 и 0.90. Эти значения обычно считают хорошими, хотя по ширине облака ситуацию вряд ли можно считать даже удовлетворительной. Аналогичные результаты получаются для других форм облака, например, эллиптической.

Поэтому мы предлагаем другой подход. Наиболее надёжной оценкой для критерия является значение его стандартного отклонения от регрессионной кривой. Регрессия строится по множеству точек $X(Y)$. Критерий с наименьшим стандартом признаётся лучшим. При этом для оценки его адекватности нужно сравнивать стандарт регрессии и стандарт экспертных оценок базы.

Результаты сравнения по разным методикам представлены в табл. 1. Мы проделали тестирование на объединённой базе для 19 наиболее распространённых критериев (см. напр. [1, 2, 9-12]). Расчёты проводились, как на полной, так и на усечённой базе. Стандартные отклонения S_K рассчитывались относительно линейных регрессий. Они используют шкалу экспертных оценок и выражаются в баллах. В таблице приведены отношения $S_K / S_Б$, а также различные коэффициенты корреляции. После каждого отношения или коэффициента курсивом указано ранжирование критерия по этой величине; первые три места выделены жирным шрифтом.

Сначала рассмотрим случай усечённой базы. По количественным значениям $S_K / S_Б$ рассмотренные критерии можно разделить на 3 группы. В первой группе содержится лишь один лидирующий критерий VIF с отношением $S_K / S_Б$ 3.52. Далее с заметным отрывом идёт очень плотная группа из 9 критериев с сотношениями от 4.05 до 4.53. В неё также входит наш критерий SGC, близкий по значениям к лидерам группы. Дальше с заметным отрывом следует третья группа с отношениями от 5.06 до 7.73. В таблице эти три группы отделены горизонтальными линиями.

Соотношение $S_K / S_Б$ позволяет не только провести ранжирование, но и оценить статистическую достоверность критериев. Очевидно, чем ближе соотношение к нулю, тем надёжнее критерий. Однако даже у лучшего критерия VIF это отношение соответствует доверительной вероятности 0.0004! Для остальных критериев доверительные вероятности ещё заметно меньше. Это показывает, что все критерии ещё очень далеки от адекватной передачи человеческого восприятия искажённого изображения. Несмотря на этот удручающий вывод, сравнение различных критериев всё же нужно производить. Наиболее надёжным мы считаем сравнение по $S_K / S_Б$. Однако рассмотрим и сравнение по коэффициентам корреляции.

Видно, что ранжирование по разным коэффициентам корреляции оказывается неодинаковым. При этом упорядочивание по коэффициенту парной корреляции Пирсона дало такой же результат что и ранжирование по $S_K / S_Б$. Нельзя утверждать, что такое совпадение будет всегда. Однако можно сделать вывод, что ранжирование по PLCC является достаточно хорошим эвристическим методом. В отличие от сравнения по стандартам, это не позволяет оценить статистическую достоверность выводов.

Упорядочивание по SRCCи KRCC для многих критериев даёт ранжирование примерно сходное с S_K. Однако некоторые критерии занимают совсем другие позиции. Тем самым, ранжирование по этим коэффициентам существенно менее надёжно, и мы не рекомендуем их использовать.

Некоторые критерии, например PSNRHA и PSNRHMA, имеют дополнительную специализацию по отдельным типам искажений. Они рассчитаны на работу с искажениями яркости или контрастности, где оказываются лучше других критериев. Поэтому интересно также тестирование критериев на полной объединённой базе. Эти результаты приведены в последних столбцах табл. 1. Все значения заметно хуже, чем в соответствующих колонках усечённой базы, что естественно. Ранжирование критериев меняется. Первое место сохраняет критерий VIF, хотя его отрыв от второго места сокращается. Критерий PSNRHA выходит с третьего места на второе. Критерий PSNRHMA уходит со второго места на четвёртое. Наш критерий SGC поднимается с шестого места на третье. Вдобавок произошла ротация нескольких критериев второй и третьей групп, а граница между ними исчезла.

Таблица 1. Тестирование критериев на объединённой базе

Критерий	Усечённая								Полная			
	$S_K/S_Б$		PLCC		SRCC		KRCC		$S_K/S_Б$		PLCC	
VIF	3,52	***1***	0,905	***1***	0,918	***2***	0,750	***1***	4,69	***1***	0,838	***1***
PSNRHMA	4,05	***2***	0,871	***2***	0,920	***1***	0,746	***2***	5,28	*4*	0,789	*4*
PSNRHA	4,06	***3***	0,871	***3***	0,913	*4*	0,735	*4*	4,94	***2***	0,818	***2***
PSNRHVSM	4,08	*4*	0,870	*4*	0,918	***3***	0,742	***3***	6,63	*12*	0,637	*12*
PSNRHVS	4,09	*5*	0,869	*5*	0,912	*5*	0,733	*5*	6,48	*11*	0,657	*11*
SGC	4,09	*6*	0,869	*6*	0,896	*7*	0,709	*7*	5,14	***3***	0,801	***3***
VSNR	4,19	*7*	0,862	*7*	0,895	*8*	0,706	*9*	7,05	*16*	0,571	*16*
WSNR	4,19	*8*	0,862	*8*	0,892	*9*	0,706	*8*	7,11	*17*	0,563	*17*
VIFP	4,29	*9*	0,854	*9*	0,873	*12*	0,682	*11*	5,89	*6*	0,729	*6*
NQM	4,53	*10*	0,836	*10*	0,844	*14*	0,646	*14*	6,64	*13*	0,636	*13*
SAC	5,06	*11*	0,791	*11*	0,848	*13*	0,646	*13*	5,95	*9*	0,722	*9*
UQI	5,24	*12*	0,773	*12*	0,764	*19*	0,565	*18*	5,91	*7*	0,726	*7*
IWSSIM	5,33	*13*	0,763	*13*	0,907	*6*	0,723	*6*	5,69	*5*	0,750	*5*
SSIM	5,36	*14*	0,761	*14*	0,883	*10*	0,686	*10*	5,92	*8*	0,725	*8*
PSNR	5,36	*15*	0,760	*15*	0,802	*17*	0,594	*17*	6,84	*14*	0,607	*14*
SNR	5,64	*16*	0,731	*16*	0,769	*18*	0,563	*19*	7,02	*15*	0,578	*15*
IFC	5,64	*17*	0,730	*17*	0,839	*15*	0,645	*15*	8,30	*19*	0,261	*19*
MSSIM	5,84	*18*	0,708	*18*	0,878	*11*	0,680	*12*	6,15	*10*	0,698	*10*
MSE	7,73	*19*	-0,353	*19*	-0,802	*16*	-0,594	*16*	8,18	*18*	-0,309	*18*

Можно дополнительно оценить стабильность работы критериев, беря сумму мест для усечённой и полной базы. Первое место занимает VIF (2 очка), на втором месте PSNRHA(5 очков), на третьем PSNRHMA (6 очков), на четвёртом SGC (9 очков), остальные имеют 15 и более очков. Это показывает, какие критерии следует предпочесть.

Заключение

Таким образом, в работе получены следующие результаты.

1. Проведён анализ двух наиболее представительных баз TID2008 [1] и LIVE [2]. Предложен метод объединения их в единую базу. Выявлено, на каких типах искажений к экспертным оценкам нужно относиться с осторожностью.

2. Показано, что сравнение критериев по коэффициентам корреляции Пирсона, Кендалла и Спирмена не информативно: эти коэффициенты могут оказываться близкими к единице, даже при явной неадекватности критериев. Для оценки критерия предложено использовать характеристику статистической достоверности: отношению стандартных уклонений базы и регрессии данного критерия.

3. На объединённой базе TID2008-LIVE проведено ранжирование всех критериев по статистической достоверности. При этом хорошее место занимает предложенный нами ранее критерий, который использует сеточную норму Соболева w_2^1. Показано, что даже лучшие критерии далеки от адекватности.

Рекомендации

1. Наилучшим оказывается критерий VIF. Его в первую очередь следует использовать в работе. Критерии второй группы также можно использовать, но с некоторой осторожностью. Критерии третьей группы существенно хуже, и мы не рекомендуем их использовать.

2. Сравнение критериев мы рекомендуем производить по $S_K / S_Б$.

3. Для сравнения критериев мы рекомендуем предложенное здесь объединение баз TID2008 (с исключением групп №12, 14-17) и LIVE.

Работа поддержана грантом РФФИ 11-01-00102.

Литература

1. Ponomarenko N., Lukin V., Zelensky A., Egiazarian K., Astola J., Carli M., Battisti F. TID2008 –A Database for Evaluationof Full-Reference Visual Quality Assessment Metrics // Успехи современной радиоэлектроники. – 2009. – №10. – С. 30–45.

2. Sheikh H.R., Seshadrinathan K., Moorthy A.K., Wang Z., Bovik A.C. and Cormack L.K. Image and video quality assessment research at LIVE // [Online] – Available: http://live.ece.utexas.edu/research/quality/.

3. Kodak Lossless True Color Image Suite // [Online] – Available: http://r0k.us/graphics/kodak/. – 7.05.2010.

4. Абрамов С.К., Зеленский А.А., Лукин В.В., Пономаренко Н.Н. Использование базы TID2008 при разработке метрик визуального качества и методов обработки изображений // Компьютерные системы и информационные технологии. – 2012. –Т.56.– №4. –С.99-109.

5. Chandler D.M. and Hemami S.S.VSNR: A wavelet-based visual signal-to-noise ratio for natural images // [Online] – Available: http://foulard.ece.cornell.edu/dmc27 /vsnr/vsnr.html.

6. Ninassi A., Le Callet P. and Autrusseau F. Subjective quality assessment—IVC database // [Online] –Available: http://www2.irccyn.ec-nantes.fr/ivcdb.

7. MICT Image Quality Evaluation Database // [Online] – Available: http://mict.eng.u-toyama.ac.jp/mictdb.html. – 01.01.2010.

8. Larson E.C. and Chandler D.M. Categorical image quality (CSIQ) database // [Online] – Available: http://vision.okstate.edu/csiq.

9. Sheikh H.R. and Bovik A.C. Image Information and Visual Quality // IEEE Transactions on Image Processing. – Vol: 15. –№: 2. – February 2006. –Page(s):430 – 444.

10. Калиткин Н.Н., Голованов Р.В., Проблема сравнения критериев оценки качества искажённого изображения // Труды конференции DSPA-2013. – Март 2013. – Москва – С.31–34.

11. Голованов Р.В. Современные методы оценки качества искажённого изображения. Базы тестовых изображений // Труды конференции МэИнфо-2013. – Апрель 2013. –Зеленоград. – С.135.

12. Калиткин Н.Н., Голованов Р.В. Критерий сглаженных градиентов для оценки качества искаженного изображения // ДАН. – Т.451. – № 4. – Август 2013. С.385-388.

STATISTICAL INVALIDATION OF ALL KNOWN IMAGE QUALITY ASSESSMENTS

Golovanov R.V., Kalitkin N.N.

A critical analysis of the most popular databases of test images TID2008and LIVE has been carried out. We propose a method of combining these databases. On the basis of new database 19 different criteria for assessing the quality of distorted images were compared. Shown that the comparison of criteria by the Pearson correlation coefficient, Spearman and Kendall is not informative.Proposed to carry outof comparison by the division of standard deviations of the image database and the criterion.Criteria were ranked according to this rule and it is shown that even the best criterion is statistically unreliable.