УДК: 004.932.4

МЕТОДИКА ОБЪЕКТИВНОЙ ОЦЕНКИ КАЧЕСТВА ВОССТАНОВЛЕНИЯ ФОНА В ВИДЕО

Боков А.А., *аспирант Московского государственного университета им. М.В. Ломоносова,*
e-mail: abokov@graphics.cs.msu.ru;
Ватолин Д.С., *к.ф.-м.н., с.н.с. Московского государственного университета им. М.В. Ломоносова,*
e-mail: dmitriy@graphics.cs.msu.ru.

OBJECTIVE QUALITY ASSESSMENT METHODOLOGY FOR VIDEO BACKGROUND RECONSTRUCTION

Bokov A.A., Vatolin D.S.

In its general form, video background reconstruction is usually defined as a task of plausible video region reconstruction that is marked with an input mask. Object removal is a typical example of background reconstruction. Several new methods were introduced over the past few years; however, no standard benchmark has yet been established. In this work we propose an objective background reconstruction quality benchmark that consists of several metrics that we demonstrate to have higher correlation with perceptual quality compared to prior approaches. Perceptual background reconstruction quality is quantitatively evaluated based on pairwise comparison of background reconstruction methods performed by over 300 human subjects.

Key words: background reconstruction, video processing, objective quality assessment, quality benchmark.

Ключевые слова: восстановление фона, обработка видео, объективная оценка качества, методика сравнения.

В общей постановке задача восстановления фона в видео состоит в максимально правдоподобной реконструкции области видеопоследовательности, отмеченной входной маской, на базе известной части видеопоследовательности, лежащей вне маски. Однако большинство авторов ограничиваются публикацией результатов работы предложенного алгоритма на нескольких видеопоследовательностях, которые зачастую различны для разных подходов. Предлагается методика для объективного сравнения алгоритмов восстановления фона в видео. Вводится ряд показателей качества восстановления фона, превосходящих предшествующие подходы по корреляции с экспертной оценкой, полученной путем попарного сравнения результатов различных алгоритмов с участием более 300 респондентов.

Введение

Одной из важных задач в области обработки видео является восстановление фона. Ее можно определить как задачу максимально правдоподобной реконструкции области видео, помеченной входной маской Ω, на базе известной части видеопоследовательности, лежащей вне маски: V/Ω.Такая постановка находит применение в ряде прикладных областей, включающем восстановление от ошибок видеокодека, возникающих при потере пакетов в процессе передачи видео по сети, восстановление архивных видеозаписей, содержащих различные дефекты пленки, удаление нежелательных объектов из видео, заполнение областей открытия в процессе конвертации видео в стереоскопический формат и многое другое.

Многие алгоритмы накладывают существенные ограничения на входные видеопоследовательности и маски областей восстановления или требуют определенные дополнительные данные для работы, ограничивая таким образом область своей применимости [27, 28] или же ориентируясь на конкретный узкий сценарий работы [4-9]. Однако ряд алгоритмов, предложенных в последние годы, позволяют решать задачу восстановления фона в достаточно широкой ее постановке. Например, алгоритм, предложенный в работе [1], позволяет работать со свободно движущейся камерой и восстанавливать динамические текстуры в видео, такие как поверхность воды, огонь, толпа людей. Алгоритм, предложенный в [2], накладывает только требования статичности восстанавливаемого фона и видимости восстанавливаемых областей хотя бы в одном кадре входного видео. Но авторы отмечают, что время работы алгоритма на одной последовательности длительностью 100 кадров достигало четырех часов на сервере с 64 логическими процессорами. Аналогичная ситуация характерна для большинства современных алгоритмов восстановления видео: время работы и потребление памяти серьезно ограничивают их применимость к длинным видеопоследовательностям высокого разрешения. В работе [3] авторы предлагают ограничить вычислительную сложность и потребление памяти за счет ограничения пространства поиска до фиксированной временной окрестности текущего кадра.

В целом, прогресс в данной области во многом ограничивается отсутствием общепризнанной методики оценки качества и сравнения различных алгоритмов восстановления фона в видео. Большинство авторов ограничиваются публикацией результатов работы алгоритма на нескольких видео, иногда вместе с результатами работы подходов-конкурентов на тех же видеопоследовательностях. Применение объективных методов

оценки качества затруднено даже при наличии эталонных результатов восстановления, так как алгоритмы восстановления видео, как правило, оцениваются по визуальному качеству восстановления с точки зрения наблюдателя, а не по близости к некоторому эталонному результату. Эта проблема приобретает особую актуальность в случаях, когда область восстановления имеет большой размер как в смысле пространственных размеров, так и временной продолжительности (например, удаление объекта из видео).

В данной работе предлагается ряд методов, позволяющих производить объективную оценку качества восстановления фона в видео, в предположении, что доступен эталонный результат восстановления. Адекватность предложенных методов оценки качества была подтверждена следующим образом:

Построены 7 тестовых видеопоследовательностей, включающих эталонный фон, который является идеальным результатом восстановления (§ 3.1).

Проведена автоматическая оценка качества результатов работы 6 алгоритмов восстановления фона на построенных видеопоследовательностях с помощью предложенных объективных показателей качества (§ 3.2).

Проведена экспертная оценка качества путем попарного сравнения результатов различных алгоритмов восстановления фона с участием более 300 респондентов (§ 3.3).

Вычислена корреляция между результатами экспертной и автоматической оценок (§ 4).

Обзор области исследований

В данном разделе представлен краткий обзор существующих подходов к оценке качества и получения эталонных результатов восстановления фона в видеопоследовательностях. Рассматриваются методы и наборы данных, которые использовались авторами алгоритмов восстановления фона для проведения объективной оценки качества и сравнения с конкурентами.

Количественная оценка качества восстановления часто используется в случаях, когда восстанавливаемый регион видео имеет относительно малый размер либо в смысле пространственных размеров, либо временной продолжительности. Это верно, например, для задачи скрытия ошибок, вызванных потерей пакетов в процессе передачи видеопоследовательности по сети [4-6]. Как правило, используются традиционные методы оценки качества видео, такие как PSNR и SSIM. Эталонный результат восстановления получается за счет того, что ошибки, эмулирующие эффект потери пакетов разной интенсивности, вносятся в исходный видеопоток искусственным образом. Авторы алгоритма, предложенного в работе [3], проводят его апробацию в ряде различных сценариев применения, однако количественная оценка качества работы проводится только для сценария скрытия ошибок, вызванных потерей пакетов при передаче видеопотока. Аналогичным образом оценка качества производится для задач автоматического устранения логотипов [7, 8] и удаления текста [9] из видеопоследовательности. Авторы работы [10] используют меру RMSE (Root-Mean-Square Error – корень из среднеквадратичной ошибки) для оценки качества восстановления достаточно крупных областей, но продолжающихся не более 5 кадров. В некоторых тестовых последовательностях целые кадры включались в область восстановления, что, фактически, сводит задачу к межкадровой интерполяции. Предложенный алгоритм восстановления был сравнен с одним альтернативным подходом на 10 последовательностях низкого разрешения (порядка 352×240) с продолжительностью от 35 до 100 кадров.

В работе [11] для построения тестовых последовательностей на каждом кадре исходной видеопоследовательности вырезается случайным образом выбранный блок. Предложенный алгоритм сравнивается с пятью альтернативными алгоритмами. Авторы работ [12] и [13] проводят сравнение на единственной последовательности, где небольшой блок вырезается из одной и той же позиции на каждом кадре. Для оценки качества используются PSNR и среднеквадратичная ошибка. В работе [14] используется простая сумма абсолютных разностей (SAD) между результатом реконструкции и оригинальной видеопоследовательностью для демонстрации высокой точности предложенного подхода. Он сравнивается с двумя альтернативными подходами на одной последовательности и трех отдельных областях восстановления, которые характеризуются различным движением (быстрое движение, медленное движение и его полное отсутствие). Однако стоит отметить, что рассматриваемые авторами области являются относительно крупными как в пространственном, так и во временном направлениях, но никакого обоснования используемой методики оценки качества не приводится. Авторы работы [15] проводят подсчет индекса SSIM на одной из тестовых последовательностей, однако приведенное сравнение с конкурирующими подходами имеет качественный, а не количественный характер. В недавнем обзоре алгоритмов восстановления фона 2014 года [16] отмечается, что авторам неизвестны работы, которые бы рассматривали проблему количественной оценки качества восстановления фона видеопоследовательностей в общей постановке.

По результатам проведенного обзора можно сделать вывод, что проблема количественной оценки качества восстановления больших пространственно-временных областей в видеопоследовательностях явным образом не рассматривалась в предшествующих работах. Классические методы оценки качества, такие как PSNR и SSIM, хорошо подходят для оценки качества восстановления областей, которые либо являются небольшими в смысле пространственных размеров, либо имеют малую продолжительность, не превышающую нескольких кадров. Однако эти методы становятся значительно менее надежными в случаях областей восстановления, которые являются крупными как в смысле пространственных размеров, так и временной продолжительности, так как соответствие результата восстановления некоторому одному эталону уже не обязательно для достижения высокого визуального качества.

Методика количественного сравнения алгоритмов восстановления фона

Построение тестового набора данных

Для количественного сравнения различных подходов, в первую очередь, требуется определение набора данных, на котором оно будет проводиться. При построении тестовых видеопоследовательностей принимался во внимание ряд принципов. Во-первых, каждая тестовая видеопоследовательность должна представлять некоторую сложность для существующих алгоритмов восстановления фона: тривиальные примеры не позволяют отделить высококачественные алгоритмы от низкокачественных. Также предпринимается попытка охватить как можно больше практических случаев восстановления фона, включая последовательности со статичной и свободно перемещающейся камерой, статический и динамический фон, динамические текстуры в видео (например, поверхность воды, огонь). Некоторые из построенных видеопоследовательностей имеют области восстановления, которые включают фрагменты фона, не видимые ни в одном из кадров входного видео, что нарушает предположение, используемое множеством алгоритмов восстановления фона (например, [2, 22]). И, наконец, все предлагаемые тестовые видеопоследовательности имеют разрешение 1920×1080 и продолжительность от 150 до 200 кадров. Это важно для выделения более практичных подходов, так как многие предлагаемые алгоритмы имеют серьезные ограничения в плане обработки продолжительных видеопоследовательностей высокого разрешения из-за низкой скорости работы и/или слишком высокого потребления памяти.

Все построенные тестовые видеопоследовательности относятся к одному способу применения алгоритмов восстановления фона – удаление объектов из видео. Для получения видеопоследовательностей с эталонным восстановленным фоном проводится искусственное наложение различных объектов переднего плана на фоновые видеопоследовательности средствами компьютерной графики. Для максимально правдоподобной интеграции объектов переднего плана в фоновое видео с учетом движения камеры используется программный пакет Blender [17]. Пример построенного видео с маской области восстановления приведен на рис. 1.

Таким образом, каждый алгоритм восстановления фона получает на вход видеопоследовательность с наложенным объектом переднего плана и соответствующей маской восстановления. Затем проводится оценка качества результата восстановления с учетом исходной видеопоследовательности, содержащей эталонные изображения фона за наложенным объектом. Всего было построено 7 тестовых видеопоследовательностей для сравнения качества различных алгоритмов восстановления фона.

а) Кадр с наложенным объектом переднего плана *б) Маска восстановления*

Рис. 1. Пример кадра из тестовой видеопоследовательности для оценки качества алгоритмов восстановления фона в видео. Целью восстановления является максимально правдоподобное удаление объекта из входного видео

а) Результат восстановления фона алгоритмом [1] *б) Результат восстановления фона алгоритмом [24]*

Рис. 2. Иллюстрация проблем традиционных методов в контексте оценки качества восстановления фона в видео. Результат слева очевидно обладает более высоким визуальным качеством, однако метрика MSE указывает на обратное

Методы оценки качества восстановления фона

В качестве исходной точки для сравнения разумно использовать традиционные для области оценки качества видео методы, такие как:

$$\mathrm{MSE}(V,V_r)=\frac{1}{|\Omega|}\sum_{x\in\Omega}\mathrm{MSE}(P(x),P_r(x)),$$

$$\mathrm{DSSIM}(V,V_r)=\frac{1}{|\Omega|}\sum_{x\in\Omega}1-\mathrm{SSIM}(P(x),P_r(x)),$$

где V и V_r обозначают видео, содержащее результат восстановления, и видео с эталонным фоном соответственно. $P(x)$ и $P_r(x)$ – это блоки результата восстановления и эталона соответственно, с центром в пикселе x и размером $n \times n \times 1$ (то есть, двумерные блоки имеющие пространственные размеры $n \times n$ и продолжительность в 1 кадр, в рамках данной работы n предполагается равным 9 пикселам). MSE – среднеквадратичная разница между соответствующими блоками, SSIM – индекс структурного сходства между блоками [18]. В рамках данной работы все показатели, основанные на использовании 2D-блоков, вычисляются для яркостной компоненты видеопоследовательности, использование цветовых компонент не приводило к значимому повышению корреляции с экспертной оценкой в проведенных экспериментах. Ω обозначает пространственно-временной регион видео, полученный из исходной области восстановления Ω_s путем ее расширения на $n/2$ пикселов в пространственных направлениях, чтобы все блоки $P(x)$, содержащие хотя бы один пиксел из Ω_s, были целиком включены в Ω: $\Omega=\{x \mid P(x)\cap\Omega_s\neq\varnothing\}$.

Однако рассмотренные выше традиционные методы обладают рядом недостатков в контексте оценки качества восстановления фона в видео (см. рис. 2). Далее данные недостатки будут рассмотрены и будут приведены возможные способы их устранения. Во-первых, внутри восстановленной области могут присутствовать сдвиги относительно эталона, к которым показатели MSE и DSSIM имеют высокую чувствительность. Наиболее простой способ устранения данного ограничения – это проведение анализа сразу на нескольких масштабах:

$$\mathrm{MSMSE}(V,V_r)=\sum_{i=0}^{M-1}w_i^{MSE}\mathrm{MSE}(V^i,V_r^i),$$

$$\mathrm{MSDSSIM}(V,V_r)=\sum_{i=0}^{M-1}w_i^{DSSIM}\mathrm{DSSIM}(V^i,V_r^i).$$

Верхний индекс i обозначает уровень Гауссовой пирамиды. То есть, V_r^0 – это исходное эталонное видео, V_r^1 – это то же самое видео, но уменьшенное в два раза по обоим пространственным направлениям (продолжительность видеопоследовательности не изменяется). M – константа, обозначающая общее количество уровней рассматриваемой Гауссовой пирамиды, w_i – веса соответствующих уровней пирамиды. Присвоение больших весов более высоким уровням пирамиды увеличивает устойчивость метода к мелким сдвигам между результатом восстановления и эталоном. Конкретные значения весов w_i вычисляются на основе проведенной экспертной оценки различных алгоритмов восстановления фона.

Человеческое зрение, как правило, более чувствительно к нестабильным во времени искажениям, однако рассмотренные до сих пор методы не пригодны для измерения такого вида искажений. Для устранения этой проблемы предлагаются следующие показатели, явным образом учитывающие стабильность во времени:

$$\mathrm{MSEdt}(V,V_r)=\frac{1}{|\Omega|}\sum_{x\in\Omega}\max\begin{pmatrix}\mathrm{MSE}(P(x),P(x+s_x))-\\-\mathrm{MSE}(P_r(x),P_r(x+s_x)),0)\end{pmatrix},$$

$$\mathrm{DSSIMdt}\left(V,V_r\right)=$$

$$=\frac{1}{|\Omega|}\sum_{x\in\Omega}\max\begin{pmatrix}\mathrm{SSIM}(P_r(x),P_r(x+s_x))-\\-\mathrm{SSIM}(P(x),P(x+s_x)),0\end{pmatrix},$$

где $s_x=(vx,vy,-1)$ обозначает вектор движения от текущего кадра к предыдущему на эталонной видеопоследовательности в пикселе x. Было опробовано несколько алгоритмов оптического потока для вычисления векторов движения на эталонной видеопоследовательности. Однако наилучший результат был получен при использовании алгоритма PatchMatch [19] с радиусом поиска ограниченным до 1/20 ширины кадра и с использованием блоков 9×9 пикселей. Такой подход обнаруживает нестабильность результата восстановления вдоль векторов движения эталона. Однако это неявно предполагает полную выравненность результата восстановления и эталона. Для смягчения этого требования вводится одновременный учет нескольких масштабов аналогично MSMSE и MSDSSIM:

$$\mathrm{MSMSEdt}\left(V,V_r\right)=\sum_{i=0}^{M-1}w_i^{MSEdt}\mathrm{MSEdt}\left(V^i,V_r^i\right),$$

$$\mathrm{MSDSSIMdt}\left(V,V_r\right)=\sum_{i=0}^{M-1}w_i^{DSSIMdt}\mathrm{DSSIMdt}\left(V^i,V_r^i\right).$$

В целом, устойчивость рассмотренных показателей к сдвигам между результатом восстановления и эталоном ограничивается лишь пространственными сдвигами небольшой величины. Для повышения устойчивости к сдвигам возможно применение более сложных методов, которые явным образом находят сдвиги, минимизирующие разницу между результатом восстановления и эталоном, причем поиск может не ограничиваться текущим кадром, а включать в себя нахождение наиболее похожего блока во всей эталонной видеопоследовательности. Более конкретно, для каждого блока результата восстановления находится наиболее похожий блок в эталонном видео и расстояния между ними суммируются:

$$C_{\#}^{3\mathrm{D}}(V,V_r)=\frac{1}{|\Omega'|}\sum_{x\in\Omega'}\min_y\#(Q(x),Q_r(y)),$$

$$C_{\#}\left(V,V_r\right)=\frac{1}{|\Omega|}\sum_{x\in\Omega}\min_y\#(P(x),P_r(y)).$$

Разница между этими двумя показателями заключается в том, что $C_{\#}^{3\mathrm{D}}$ использует пространственно-временные 3D-блоки, обозначаемые как $Q(x)$ (используются 5×5×5 RGB блоки в соответствии с [1]), а не традиционные 2D-блоки $P(x)$. Ω', соответственно, обозначает область восстановления, расширенную на 3 пик-

села во всех направлениях, включая временное (это гарантирует включение всех 3D-блоков, содержащих хотя бы 1 пиксел области восстановления). # – это функция, которая используется для вычисления степени сходства блоков (или, другими словами, расстояния между блоками).

Использование пространственно-временных блоков позволяет также учитывать сходство характера движения сравниваемых областей видеопоследовательности. Авторы работы [1] предлагают модифицированный метод измерения сходства пространственно-временных блоков, который позволил улучшить визуальное качество восстановленного фона в рамках предложенного ими алгоритма:

$$\mathrm{TMSE}(Q(x), Q_r(y)) =$$
$$= \frac{1}{N}(\| Q(x) - Q_r(y) \|_2^2 + \lambda \| T(x) - T_r(y) \|_2^2),$$

где $T(x)$ и $T_r(y)$ – это 3D-блоки текстурных признаков, определенных в [1], из результата восстановления и эталона соответственно, N – количество пикселов в блоке .TMSE, наряду с простой среднеквадратичной разницей MSE, используется как мера сходства блоков в $C_{\#}^{3\mathrm{D}}$. MSE и DSSIM используются как меры сходства блоков в $C_{\#}$.

Таким образом, можно определить многомасштабные показатели $C_{\mathrm{MSE}}^{\mathrm{MS3D}}$, $C_{\mathrm{TMSE}}^{\mathrm{MS3D}}$, $C_{\mathrm{MSE}}^{\mathrm{MS}}$, $C_{\mathrm{DSSIM}}^{\mathrm{MS}}$ аналогично определению MSMSE. Использование результатов на разных масштабах в рамках подходов, основанных на нахождении наиболее похожего блока во всем эталонном видео, позволяет отдельно учитывать искажения разного масштаба, которые, как правило, имеют разную заметность для зрительной системы человека [20].

В то время как $C_{\#}^{3\mathrm{D}}$ позволяет учитывать сходство движения между результатом восстановления и эталоном за счет использования 3D-блоков, $C_{\#}$ учитывает лишь пространственные искажения. Для устранения этого ограничения предлагается измерить, насколько расстояние до наиболее похожего блока в эталонном видео изменяется от кадра к кадру. Более конкретно, для каждого блока на текущем кадре результата восстановления фона находится наиболее похожий блок на предыдущем кадре в рамках окна ограниченных размеров, а затем проводится сравнение расстояния до наиболее похожего блока эталона между ними:

$$\mathrm{C}_{\#\mathrm{dt}}(V, V_r) =$$
$$= \frac{1}{|\Omega|}\sum_{x\in\Omega}\left|\min_y(\#(P(x), P_r(y))) - \min_y(\#(P(x_{prev}), P_r(y)))\right|,$$
$$x_{prev} = \underset{y\in\Omega_{prev}^{w\times w}(x)}{\operatorname{argmin}} \#(P(x), P(y)), \# = \mathrm{DSSIM},\ \mathrm{MSE}.$$

$\Omega_{prev}^{w\times w}(x)$ – это окно размера $w\times w$ пикселов (в данной работе используется w равное 1/10 ширины кадра), которое пространственно центрировано по пикселу x, но находится на предыдущем кадре по отношению к нему. Версии, учитывающие результаты на разных масштабах, $\mathrm{C}_{\mathrm{MSEdt}}^{\mathrm{MS}}$ и $\mathrm{C}_{\mathrm{DSSIMdt}}^{\mathrm{MS}}$ определяются аналогично MSMSEdt. В отличие от показателя MSMSEdt, который позволяет оценивать стабильность результата восстановления вдоль векторов движения эталонного видео (это предполагает необходимость полного соответствия движения в результате восстановления и движения в эталоне, что может быть слишком сильным предположением в ряде случаев), данные оценки измеряют то, насколько искажения блоков результата восстановления варьируются от кадра к кадру.

Точное вычисление показателей, основанных на нахождении наиболее похожего блока в эталонном видео, затруднено чрезмерно высокой вычислительной сложностью полного перебора всех блоков эталонной видеопоследовательности. Гораздо более эффективным образом можно вычислить приблизительные оценки с помощью алгоритма PatchMatch [19]. В случае 3D-блоков данный алгоритм применяется в полной аналогии с работой [1]. Для нахождения наиболее похожих 2D-блоков алгоритм PatchMatch применяется в покадровом режиме, результат предыдущего кадра используется как начальное приближение в текущем для ускорения сходимости. Также, для сравнения блоков всегда используется метрика SSD (сумма квадратов разностей) в цветовом пространстве RGB. Явная минимизация индекса SSIM в проведенных экспериментах не приводила к существенному повышению корреляции с результатами экспертной оценки, в то время как ее использование многократно увеличивает вычислительную сложность алгоритма оценки качества восстановления.

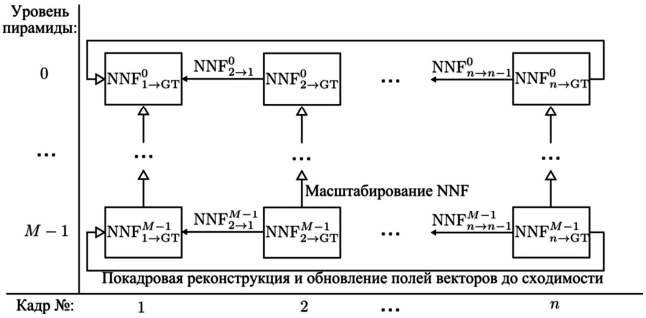


Рис. 3. Общая схема простого жадного алгоритма, используемого для совместной минимизации показателей C_{MSE}^{MS} и C_{MSEdt}^{MS} на базе предоставленного начального приближения. В ходе алгоритма итеративно вычисляются межкадровые поля векторов $NNF_{t\to(t-1)}^{M-1}$, соединяющие наиболее похожие блоки кадров с номерами t и $t-1$ на нижнем $(M-1)$-ом уровне Гауссовой пирамиды, и поля векторов до наиболее похожих блоков эталона $NNF_{t\to GT}^{M-1}$ для каждого кадра с помощью алгоритма PatchMatch [19]. На их основе производится покадровая реконструкция. Результат реконструкции используется для вычисления обновленных полей векторов, которые, в свою очередь, используются для проведения следующей итерации реконструкции. После окончания обработки текущего уровня пирамиды происходит переход на следующий уровень, где процесс повторяется

Проведение экспертной оценки качества

Немногие работы в области восстановления фона в видео предоставляют исходный код предложенных алгоритмов. Поэтому для увеличения количества данных для проведения экспертной оценки качества работы различных алгоритмов и измерения корреляции пред-

ложенных объективных показателей с визуальным качеством в данной работе также оцениваются коммерческие инструменты восстановления фона в видео и алгоритмы восстановления фона в изображениях. Итого в сравнении приняло участие 6 алгоритмов:

- Video Inpainting of Complex Scenes [1].
- Алгоритм BGR, описанный в работе [22].
- Инструмент F_RigRemoval из программного пакета Nuke [21].
- Инструмент Remove Rig из программного пакета PFClean [23].
- Простой алгоритм восстановления для изображений Telea Inpainting [24].
- Более сложный алгоритм восстановления для изображений Image Completion using Planar Structure Guidance [25].

Для дополнительного расширения используемого набора алгоритмов также добавляется ряд синтетических результатов, получаемых путем прямой совместной минимизации показателей C^{MS}_{MSE} и C^{MS}_{MSEdt}. Это является дополнительным способом проверки того, что более низкие значения C^{MS}_{MSE} и C^{MS}_{MSEdt} соответствуют более высокому визуальному качеству. Это не очевидно, так как подход, на котором основаны эти показатели, явным образом не использует сходство движения между результатом восстановления и эталоном. Для достижения такой минимизации используется модификация алгоритма реконструкции, предложенного в работе [1]. Общая схема алгоритма представлена на рис. 3. Основная модификация заключается в использовании двух полей векторов в процессе реконструкции – поля $\mathrm{NNF}_{t\to(t-1)}$, которое соединяет наиболее похожие блоки между текущим и предыдущим кадрами результата реконструкции, и поля $\mathrm{NNF}_{t\to GT}$, которое соединяет блоки текущего кадра результата реконструкции с наиболее похожими блоками во всем эталонном видео. Тогда, одна итерация реконструкции пиксела x на текущем кадре результата восстановления V имеет следующий вид:

$$V[x]=\left(\frac{(1-\tau)\sum_{y\in P(x)}w^s(y)V_r[x+\mathrm{NNF}_{t\to GT}[y]]+}{+\tau\sum_{y\in P(x)}w^t(y)V[x+\mathrm{NNF}_{t\to(t-1)}[y]]}{(1-\tau)\sum_{y\in P(x)}w^s(y)+\tau\sum_{y\in P(x)}w^t(y)}\right),$$

где τ – это константа, обозначающая вес темпоральной компоненты, которая использует результат реконструкции предыдущего кадра для реконструкции текущего (в данной работе используется $\tau=0,4$). $w^s(y)$ и $w^t(y)$ обозначают весовые функции, которые присваивают большие веса тем векторам полей $\mathrm{NNF}_{t\to GT}$ и $\mathrm{NNF}_{t\to(t-1)}$, соответственно, которые соединяют более похожие блоки (с меньшим значением функции расстояния SSD между блоками). Используются экспоненциальные весовые функции по аналогии с работой [1].

Такой подход позволяет находить локальные оптимумы, которые могут зависеть от начального приближения. Результаты всех 6 алгоритмов использовались в качестве начальных приближений: во всех случаях предложенный подход позволил существенно уменьшить значения обоих показателей (рис. 4). Добавление таким образом полученных синтетических результатов удваивает количество доступных данных.

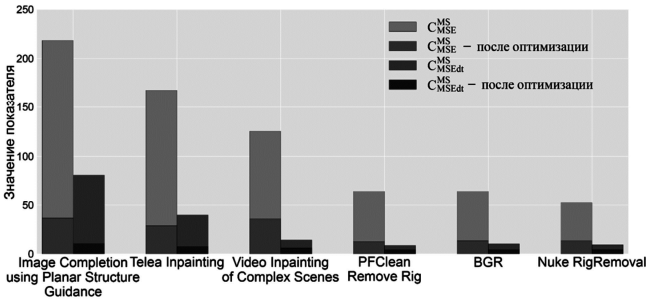


Рис. 4. Значения показателей C^{MS}_{MSE} и C^{MS}_{MSEdt} до и после предложенного алгоритма оптимизации, усредненные по всем тестовым видеопоследовательностям

Для количественной оценки визуального качества было проведено исследование, в рамках которого участники попарно сравнивали результаты различных алгоритмов и выбирали результат с наилучшим, по их мнению, качеством. Каждый участник проводил сравнение 28 пар результатов восстановления фона, которые включали 3 контрольных вопроса, где требовалось сравнить эталонный результат с результатом заведомо низкокачественного алгоритма. Для успешного прохождения требовалось ответить на все вопросы, включая правильные ответы на все контрольные вопросы. Общее количество пар, требующих сравнения, составило 2964. Это число составлено из 6 оригинальных алгоритмов, 6 соответствующих синтетических алгоритмов, состоящих в применении вышеописанной процедуры совместной минимизации показателей C^{MS}_{MSE} и C^{MS}_{MSEdt} к результатам оригинальных алгоритмов, и 19 видеопоследовательностей, на которых производилось их сравнение (использовались различные фрагменты фиксированного размера из исходных 7 видео). Итого было собрано 8533 попарных сравнений, произведенных 341 участником, которые затем были преобразованы в субъективные ранги с помощью модели Тёрстоуна [26] как для каждой последовательности по отдельности, так и для всех последовательностей сразу (рис. 5).

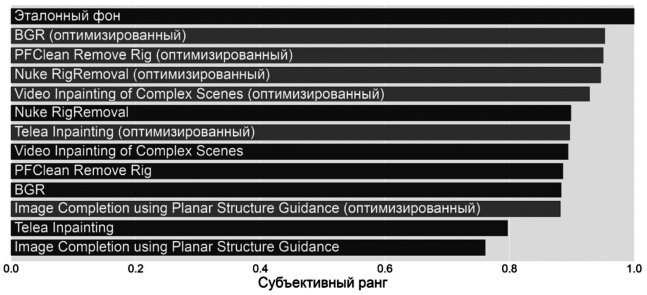


Рис. 5. Общее ранжирование протестированных алгоритмов восстановления фона в видео, вычисленное на базе попарных сравнений, проведенных более чем 300 респондентами

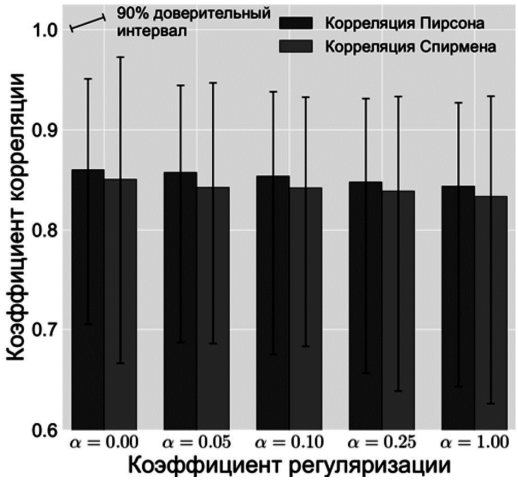
Результаты исследований

Для получения оптимальных весов в показателях, включающих результаты сразу на нескольких масшта-

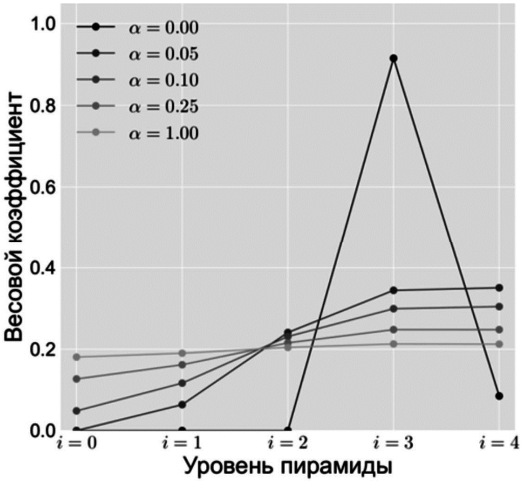
бах, производится максимизация корреляции с субъективными рангами. Сумма весов фиксируется равной единице, и используется регуляризация Тихонова:

$$\boldsymbol{w} = \underset{\boldsymbol{w}=[w_0,..,w_4]|\sum w_i=1}{\operatorname{argmax}} \sum_{s=0}^{S} \operatorname{corr}(-\log(\boldsymbol{w}\boldsymbol{M}_s), \boldsymbol{r}_s) - \alpha \parallel \boldsymbol{w}_2 \parallel,$$

где $\boldsymbol{M}_s = \{m_{ij}^s\}, m_{ij}^s$ – значение показателя для i -ого уровня пирамиды, j -ого алгоритма и s -ой видеопоследовательности (S – общее количество последовательностей), $\boldsymbol{r}_s$ – соответствующий набор субъективных рангов. Эффект регуляризации проиллюстрирован на рис. 6.



а) Корреляция показателя MSDSSIM *с субъективными рангами*



б) Распределение весов между разными уровнями пирамиды (масштабами) для показателя MSDSSIM

Рис. 6. Иллюстрация эффекта регуляризации при выборе оптимальных весов в ходе максимизации корреляции с субъективными рангами

Она повышает надежность соответствующих показателей за счет покрытия большего числа масштабов, но ценой небольшой потери в значениях корреляции с субъективными рангами. Более корректным способом выбора весов является явное использование синтетических примеров с искажениями на разных масштабах, как это, например, делается авторами метода MSSSIM [20]. Однако такой подход является слишком трудозатратным в рамках рассматриваемой задачи, поэтому итоговые веса получаются простой максимизацией корреляции на всех имеющихся данных с коэффициентом регуляризации $\alpha = 0{,}1$. Результирующие значения корреляции показаны на рис. 7.

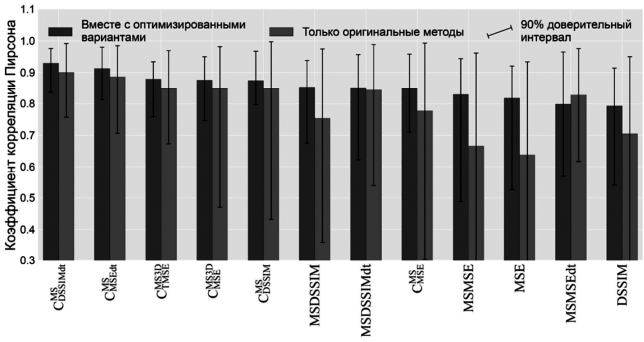


Рис. 7. Итоговые корреляции рассмотренных объективных показателей качества восстановления фона с субъективными рангами

Оценка корреляции производилась на двух различных наборах данных: полном наборе, включающем синтетические результаты, полученные путем предложенной процедуры оптимизации, и наборе, включающем только результаты 6 оригинальных алгоритмов восстановления фона.

В первую очередь, стоит отметить, что практически все объективные методы оценки качества показывают худший результат на наборе данных, включающем только 6 оригинальных алгоритмов, особенно показатели, которые базируются на прямом сравнении с эталоном (MSE, DSSIM, MSMSE, MSDSSIM). Это можно объяснить тем, что предложенная процедура оптимизации уменьшает различия между результатом восстановления и эталоном (что неудивительно, т.к. эталонный результат в процессе оптимизации напрямую используется) и в то же время увеличивает субъективные ранги (рис. 4). Сочетание этих двух факторов естественным образом приводит к повышению корреляции таких простых показателей как MSE и DSSIM на полном наборе данных. Независимо от используемого набора данных подходы, которые базируются на индексе структурного сходства (SSIM), стабильно показывают лучший результат по сравнению с подходами, использующими среднеквадратичную ошибку (MSE). Также, оценки временной стабильности, как правило, показывают более высокую корреляцию с субъективными рангами по сравнению с оценками пространственной ошибки, что и следовало ожидать. С другой стороны, показатели $\mathrm{C}^{\mathrm{MS}}_{\mathrm{MSEdt}}$ и $\mathrm{C}^{\mathrm{MS}}_{\mathrm{DSSIMdt}}$ показывают на удивление высокий результат, учитывая, что явным образом сходство движения между результатом восстановления и эталоном в них не используется. Оценки, использующие 3D-блоки, также показывают неплохую корреляцию с субъективными рангами, но они имеют более высокую вычислительную сложность и большое потребление памяти по сравнению с подходами, основанными на использовании 2D-блоков. Таким образом, следующие показатели были выбраны для объективной оценки качества восстановления фона в видео:

$$\mathrm{C}^{\mathrm{MS}}_{\mathrm{DSSIMdt}}, \left\{w_i^{CDSSIMdt}\right\} = [0.00; 0.08; 0.25; 0.30; 0.37],$$

$$\mathrm{C}^{\mathrm{MS}}_{\mathrm{DSSIM}}, \left\{w_i^{CDSSIM}\right\} = [0.04; 0.11; 0.21; 0.29; 0.35],$$

$$\mathrm{MSDSSIMdt}, \left\{w_i^{DSSIMdt}\right\} = [0.00; 0.00; 0.30; 0.32; 0.38],$$

$\mathrm{MSDSSIM}, \left\{ w_i^{DSSIM} \right\} = [0.05; 0.12; 0.23; 0.30; 0.30]$.

Заключение

В данной работе представлена методика объективной оценки качества восстановления фона в видеопоследовательностях, состоящая из четырех показателей, которые превзошли традиционные подходы к оценке качества видео. Это продемонстрировано на результатах экспертной оценки 6 алгоритмов восстановления фона на 7 построенных тестовых видеопоследовательностях с эталонным фоном, проведенной с участием более 300 респондентов. Предложенные показатели позволяют проводить объективное сравнение различных алгоритмов восстановления фона даже в случаях крупных и продолжительных по времени областей восстановления (например, удаление объекта из видео), что представляло серьезные проблемы для предшествующих методик количественного сравнения качества восстановления. Предложенные объективные оценки существенно различаются по корреляции с результатами экспертной оценки, однако имеют отдельные интуитивные интерпретации и позволяют более точно оценить слабые и сильные стороны того или иного алгоритма восстановления фона в видео.

Исследование выполнено при поддержке РФФИ в рамках научного проекта 15-01-08632 а.

Литература

1. Newson A., Almansa A., Fradet M., Gousseau Y. and Pérez P., «Video inpainting of complex scenes», SIAM Journal on Imaging Sciences, pp. 1993–2019, 2014.

2. Granados M., Tompkin J., Kim K., Grau O., Kautz J., and C. Theobalt, «How not to be seen-object removal from videos of crowded scenes», Computer Graphics Forum, volume 31, – pp. 219–228, 2012.

3. Ebdelli M., Meur O. Le and Guillemot C. «Video inpainting with short-term windows: application to object removal and error concealment», IEEE Transactions on Image Processing, – pp. 3034–3047, 2015.

4. Chen Y., Hu Y., Au O.C., Li H. and Chen C.W. «Video error concealment using spatio-temporal boundary matching and partial differential equation», IEEE Transactions on Multimedia, pp. 2-15, 2008.

5. Koloda J., Ostergaard J., Jensen S.H., Peinado A.M. and Sanchez V., «Sequential error concealment for video/images by weighted template matching», IEEE Data Compression Conference, – pp. 159–168, 2012.

6. Koloda J., Ostergaard J., Jensen S.H., Sanchez V. and Peinado A.M. «Sequential error concealment for video/images by sparse linear prediction», IEEE Transactions on Multimedia, pp. 957–969, 2013.

7. Erofeev M and Vatolin D. «Automatic logo removal for semitransparent and animated logos», Proceedings of GraphiCon 2011, pp. 26–30, 2011.

8. Yan W.Q., Wang J. and Kankanhalli M.S. «Automatic video logo detection and removal», Multimedia Systems, pp. 379-391, – 2005.

9. Mosleh A., Bouguila N. and Hamza A.B. «Automatic inpainting scheme for video text detection and removal», IEEE Transactions on Image Processing, pp. 4460-4472, 2013.

10. Shiratori T., Matsushita Y., Tang X. and Kang S.B. «Video completion by motion field transfer», IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 411–418, 2006.

11. Hu W., Tao D., Zhang W., Xie Y. and Yan Y. «A new low-rank tensor model for video completion», arXiv preprint arXiv:1509.02027, 2015.

12. Mosleh A., Bouguila N. and Hamza A.B. «Video completion using bandlet transform», IEEE Transactions on Multimedia, pp. 1591–1601, 2012.

13. Mosleh A., Bouguila N. and Hamza A.B. «Bandlet-based sparsity regularization in video inpainting», Journal of Visual Communication and Image Representation, pp. 855–863, 2014.

14. You S., Tan R.T., Kawakami R. and Ikeuchi K. «Robust and fast motion estimation for video completion», IAPR International Conference on Machine Vision Applications (MVA), – pp. 181–184, 2013.

15. Benoit J. and Paquette E. «Localized search for high definition video completion», Journal of WSCG, pp. 45–54, 2015.

16. Ilan S. and Shamir A. «A survey on data-driven video completion», Computer Graphics Forum, pp. 60–85, 2015.

17. Blender. https://www.blender.org/.

18. Wang Z., Bovik A.C., Sheikh H.R. and Simoncelli E.P. «mage quality assessment: from error visibility to structural similarity», IEEE Transactions on Image Processing, pp. 600–612, 2004.

19. Barnes C., Shechtman E., Finkelstein A. and Goldman D. «Patchmatch: A randomized correspondence algorithm for structural image editing», ACM Transactions on Graphics (TOG), 2009.

20. Wang Z., Simoncelli E.P. and Bovik A.C. «Multiscale structural similarity for image quality assessment», Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, pp. 1398–1402, 2003.

21. The Foundry Nuke. https://www.thefoundry.co.uk/products/nuke/.

22. Зачесов А.А., Ерофеев М.В., Ватолин Д.С. «Использование карт глубины при восстановлении фона в видеопоследовательностях», Новые информационные технологии в автоматизированных системах: материалы научно-практического семинара, 2015.

23. Pixel Farm PFClean. http://www.thepixelfarm.co.uk/pfclean/.

24. Telea A. «An image inpainting technique based on the fast marching method», Journal of graphics tools, pp. 23–34, 2004.

25. Huang J.-B., Kang S.B., Ahuja N. and Kopf J. «Image completion using planar structure guidance», ACM Transactions on Graphics (TOG), 2014.

26. Thurstone L.L. «A law of comparative judgment», Psychological review, 1927.

27. Herling J. and Broll W. «High-quality real-time video inpainting with PixMix», IEEE Transactions on Visualization and Computer Graphics, pp. 866–879, 2014.

28. Patwardhan K. and Sapiro G. «Video inpainting under constrained camera motion», IEEE Transactions on Image Processing, pp. 545–553, 2007.