Signal, Image and Video Processing
https://doi.org/10.1007/s11760-018-1387-5

ORIGINAL PAPER

Toward an objective benchmark for video completion

Alexander Bokov[1] · Dmitriy Vatolin[1] · Mikhail Erofeev[1] · Yury Gitman[1]

Received: 12 November 2016 / Revised: 15 October 2018 / Accepted: 1 November 2018

Abstract
Video-completion methods aim to complete selected regions of a video sequence in a natural looking manner with little to no additional user interaction. Numerous algorithms were proposed to solve this problem; however, a unified benchmark to quantify the progress in the field is still lacking. Video-completion results are usually judged by their plausibility and aren't expected to adhere to one ground-truth result, which complicates measuring the video-completion performance. In this paper, we address this problem by proposing a set of full-reference quality metrics that outperform naïve approaches and an online benchmark for video-completion algorithms. We construct seven test sequences with ground-truth video-completion results by composing various foreground objects over a set of background videos. Using this dataset, we conduct an extensive comparative study of video-completion perceptual quality involving six algorithms and over 300 human participants. Finally, we show that by relaxing the requirement of complete adherence to ground truth and by taking into account temporal consistency we can increase the correlation of objective quality metrics with perceptual completion quality on the proposed dataset.

Keywords Video completion · Inpainting · Performance evaluation

1 Introduction

Video completion is a long-standing problem in video processing and has a wide variety of real-life applications, including video restoration, rig removal and occlusion filling in virtual-view synthesis. Ilan and Shamir [14] identify subtypes of the general video-completion problem that depends on the assumed constraints for the input video as well as additional input data that some methods may require. In this paper, we assume no particular constraints on the motion of the camera or objects in a scene, and we assume that no additional input data is available besides the completion mask (i.e., we consider the general video-completion problem). Most proposed methods are far from being able to solve such a general problem; they also have high computation time and memory requirements, which become prohibitive in practical cases of long high-resolution sequences. A notable exception appears in [20]; this approach can handle a wide variety of cases in the same framework and compares favorably with prior approaches in completion quality and computational complexity. The method proposed by Granados et al. [11] can also handle many real-world cases while assuming only that the background is static and that each occluded background fragment is visible in at least one frame. The authors present results for high-resolution sequences, but they note that completion of one sequence could take up to four hours on a mainframe with 64 processors. In [8], Ebdelli et al. propose reducing the computation time by restricting the search space to a limited temporal neighborhood of the current frame.

We believe progress in this field is somewhat impeded by the absence of a benchmark to quantify advancements. As Ilan and Shamir [14] note, very few works go beyond publishing the output videos. In their survey [14], they only discuss methods of objective quality assessment for image inpainting, noting that they found no works that consider quantitative assessment of video-completion quality. This situation complicates attempts to compare existing approaches and to identify state-of-the-art methods. One way to overcome this problem is to establish a standard video-completion benchmark that includes a diverse set of challenging real-world examples and a perceptually motivated metric to evaluate

This study was funded by the RFBR under research project 15-01-08632 A.

✉ Alexander Bokov
abokov@graphics.cs.msu.ru

[1] Graphics and Media Lab, Lomonosov Moscow State University, Moscow, Russia 119991

Published online: 09 November 2018

the performance of different approaches. Even when ground-truth results are available, however, objective assessment of perceptual quality can be very difficult, as video-completion results are seldom explicitly expected to adhere to ground truth and are judged only by their plausibility, which is assessed by a human observer. This problem becomes even more relevant in cases that require filling of large spatiotemporal holes (e.g., removal of objects from the video).

Our main contributions include a challenging data set that consists of seven video sequences with ground-truth completion results as well as several full-reference objective quality metrics, which we selected using a subjective comparison of video-completion methods by over 300 human subjects. In addition, we offer an online benchmark that enables both visual and objective comparison of video-completion methods using the proposed metrics; this benchmark is available at http://videocompletion.org.

2 Related work

Most existing digital image and video quality assessment metrics [22,28] are tailored to distortions that commonly occur in a typical video acquisition and transmission pipeline, including noise, blur, compression artifacts, such as blocking and ringing. Some methods consider quality assessment of computer-generated imagery [5] or try to evaluate image illumination quality [1]. However, artifacts produced by video-completion methods can be quite different, so in this section we specifically focus on existing quality-evaluation and ground-truth acquisition methods for video completion.

Quantitative quality assessment is commonly used when the region that needs filling is relatively small in either temporal or spatial dimensions—which is definitely the case for errors that result from packet loss in video transmission [6,15,16]. Traditional quality metrics like PSNR and SSIM [28] often serve to evaluate the error concealment result relative to the original undamaged video for different packet-loss rates. Ebdelli et al. [8] apply their proposed inpainting method to a variety of scenarios, but they perform objective quality assessment only for the case of error concealment. Similarly, researchers have applied objective quality assessment to automatic logo removal [10,31] and text removal [18] from video. Shiratori et al. [23] assess root-mean-square (RMS) error when completing spatially large holes that span 1–5 frames. Entire frames were missing from some tested sequences, essentially making the task equivalent to frame interpolation. The authors compared the proposed algorithm with an alternative approach for 10 short low-resolution sequences ranging from 35 to 100 frames. Mosleh et al. in [17,19] perform objective comparison for a single sequence in which a small patch is removed from the same location in each frame. They used the PSNR and MSE metrics to compare the completion result against the original undamaged sequence. You et al. [32] employ a straightforward sum of absolute differences (SAD) between the reconstructed and original undamaged versions of the sequence to quantitatively demonstrate the accuracy of their proposed approach. Notably, the damaged regions are relatively large both spatially and temporally, but the authors provide no justification for their proposed quality assessment method. Benoit and Paquette [3] compute SSIM values for one of the test sequences but provide only a qualitative comparison with other approaches.

To summarize, prior work has not explicitly addressed the problem of video-completion quality assessment for larger spatiotemporal holes. Traditional quality metrics like PSNR and SSIM are a good fit when the damaged region is small either temporally or spatially, but they become decreasingly robust in cases where the hole is large in both spatial and temporal dimensions, as complete adherence to ground truth can no longer be expected.

3 Benchmark

3.1 Data set

We employ several guiding principles when constructing test sequences for the benchmark. First, each sequence should present at least some kind of challenge for existing methods; therefore, we avoid trivial examples. We also attempt to cover as many distinct video-completion cases as possible, including both static and free-moving cameras, static/dynamic backgrounds, and stochastic video textures. Finally, all sequences are in Full HD resolution and range from 150 to 200 frames in length. It's important to encourage more-practical approaches, as many existing methods are limited to short and low-resolution sequences owing to the processing-time and memory requirements.

For this benchmark, we consider object removal. Therefore, to obtain test sequences with ground-truth completion results, we compose various foreground objects over a set of background videos. Some of these background videos include left-view sequences from the stereoscopic-video data set RMIT3dv [7]. As foreground objects, we use those employed in the videomatting benchmark [9] as well as several 3D models. To seamlessly insert a 3D model in a background video, we use Blender motion-tracking tools [4]. Each video-completion method takes the composited sequence and the corresponding object mask as input. The benchmark then evaluates the completion results using the original background video, which serves as ground truth. Figure 1 presents thumbnails from seven proposed sequences.

Fig. 1 Thumbnails of the proposed benchmark sequences. The objective is to remove the objects highlighted in red from each video as seamlessly and plausibly as possible

3.2 Objective metrics

Standard image-quality metrics such as MSE and SSIM serve as a sensible baseline. Because working with distance metrics is more convenient, we define the following:

$$\mathrm{DSSIM}(V, V_{\mathrm{ref}}) = \frac{1}{|\Omega|} \sum_{x \in \Omega} 1 - \mathrm{SSIM}\left(P(x), P_{\mathrm{ref}}(x)\right), \quad (1)$$

$$\mathrm{MSE}(V, V_{\mathrm{ref}}) = \frac{1}{|\Omega|} \sum_{x \in \Omega} \mathrm{MSE}\left(P(x), P_{\mathrm{ref}}(x)\right). \quad (2)$$

V and V_{ref} are the inpainting result and ground-truth video, respectively. $P(x)$ and $P_{\mathrm{ref}}(x)$ are spatial 2D patches, centered at x, taken, respectively, from the result and ground truth. To define MSE and SSIM between image patches, we refer to [28]. We use $9 \times 9 \times 1$ patches of luminance values throughout the paper (taking into account the failure of chroma channels to produce noticeable improvement in our experiments). Here, $\Omega = \{x | P(x) \cap \Omega_s \neq \emptyset\}$—that is, Ω is a spatially dilated version of the input spatiotemporal hole Ω_s.

In general, however, the traditional metrics (1) and (2) are poorly suited to assessing video-completion quality, owing to a number of limitations (see Fig. 2). But several approaches can overcome these limitations. First, the completion result need not be well aligned with ground truth to have high perceptual quality, whereas both MSE and DSSIM are highly sensitive to shifts. Probably the simplest way to address this issue is to use multiscale metrics, i.e., metrics that integrate results from several scales within a Gaussian pyramid:

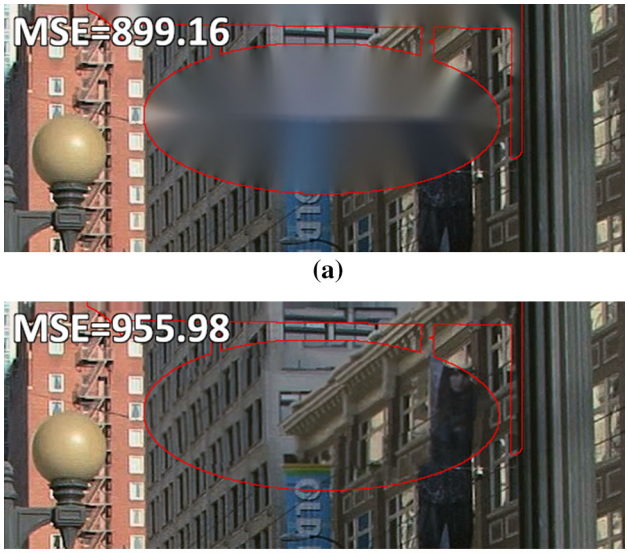

Fig. 2 Illustration of the MSE metric's limitations in the context of full-reference video-inpainting quality assessment. The lower result clearly has higher perceptual quality, but it's worse according to MSE. **a** Completion result using [25]. **b** Completion result using [20]

$$\mathrm{MS}-\#(V, V_{\mathrm{ref}}) = \sum_{i=0}^{M-1} w_i^{\#} \cdot \#(V^i, V_{\mathrm{ref}}^i), \quad \# = \mathrm{DSSIM}, \mathrm{MSE} \quad (3)$$

We use # as a placeholder for the name of the underlying metric (e.g., DSSIM or MSE in this case) throughout the paper to make the notation more concise. The superscript i denotes the level of the Gaussian pyramid—that is, V_{ref}^0 is the original ground-truth video, and V_{ref}^1 is the video blurred and subsampled by a factor of two in both spatial dimensions. M is a constant that determines the total number of levels in the pyramid; $w_i^{\#}$ are the weights of the respective pyramid levels (or scales). By assigning higher weights to higher levels, the metric becomes more shift invariant and therefore more robust to fine-scale misalignments between the completion result and ground truth. The exact values of these weights are based on the subjective-evaluation data (see Sect. 4).

Human vision may exhibit more or less sensitivity to temporal incoherence than to temporally stable spatial errors. The baseline metrics, however, are unable to distinguish between these types of errors. To overcome this problem, we propose the following metrics that explicitly capture the temporal instability:

$$\#\mathrm{d}t(V, V_{\mathrm{ref}}) = \frac{1}{|\Omega|}\sum_{x\in\Omega}\max\Big(\#\left(P(x), P(x+s_x)\right) - \#\left(P_{\mathrm{ref}}(x), P_{ref}(x+s_x)\right), 0\Big) \tag{4}$$

$$\mathrm{MS}-\#\mathrm{d}t(V, V_{\mathrm{ref}}) = \sum_{i=0}^{M-1} w_i^{\#\mathrm{d}t}\cdot\#\mathrm{d}t(V^i, V_{\mathrm{ref}}^i), \quad \# = \mathrm{DSSIM}, \mathrm{MSE} \tag{5}$$

Here, $s_x = (vx, vy, -1)$ is the optical-flow vector from the current frame to the previous one in the ground-truth sequence. Metric (5) captures the consistency of the inpainting results along ground-truth motion vectors. It assumes that completion with high perceptual quality should exhibit motion very similar to ground-truth motion, although this assumption may be too strong in some cases. We tried several optical-flow algorithms, but PatchMatch [2] with a limited search radius tends to perform best (we use 1/20th of the frame width as the search radius and use the same patch size as the metric).

The tolerance of metrics (3) and (5) to shift between the completion result and ground truth remains fairly low, however, despite the use of a multiscale scheme. To make the metrics more shift invariant, we consider a commonly used coherence measure, which some image- and video-completion algorithms [12,20,30] explicitly optimize for. It's based on finding the most similar patch in the known region for each patch of the inpainting result, assuming the plausible completion is the one that is locally similar to known image or video regions. For video completion, this measure typically uses 3D spatiotemporal patches to also take motion consistency into account [20,30]. The coherence measure is easily adaptable to the full-reference setting. For each patch of the completion result, we find the most similar patch from the whole ground-truth video and sum all the distances between them:

$$\mathrm{C}_{\#}^{\mathrm{3D}}(V, V_{\mathrm{ref}}) = \frac{1}{|\Omega^{'}|}\sum_{x\in\Omega^{'}}\min_y \#\left(Q(x), Q_{\mathrm{ref}}(y)\right) \tag{6}$$

$$\mathrm{C}_{\#}(V, V_{\mathrm{ref}}) = \frac{1}{|\Omega|}\sum_{x\in\Omega}\min_y \#\left(P(x), P_{\mathrm{ref}}(y)\right) \tag{7}$$

The only difference between these two measures is that $\mathrm{C}_{\#}^{\mathrm{3D}}$ employs 3D spatiotemporal patches, denoted as $Q(x)$ (we use $5\times5\times5$ RGB patches as [20] suggests), as opposed to 2D patches, denoted as $P(x)$. Also, $\Omega^{'}$ is the spatiotemporal hole dilated in all three dimensions by three pixels. # denotes a metric that computes the distance between patches.

Newson et al. [20] have proposed an improved distance metric for 3D spatiotemporal patches that yields better completion quality, especially in the case of stochastic textures (waves, fire, etc.). We use it as an additional metric in (6) alongside MSE, as it should naturally perform better on quality assessment too. We define it as follows:

$$\mathrm{T}-\mathrm{MSE}\left(Q(x), Q_{\mathrm{ref}}(y)\right) = \frac{1}{N}\big(||Q(x)-Q_{\mathrm{ref}}(y)||_2^2 + \lambda||T(x)-T_{\mathrm{ref}}(y)||_2^2\big) \tag{8}$$

$T(x)$ and $T_{\mathrm{ref}}(y)$ are 3D patches of texture features (as defined in [20]) from the inpainting result and ground truth, respectively. N is the number of pixels in a patch. This patch-distance metric extends to different levels of the Gaussian pyramid in accordance with [20]. We keep the multiscale scheme, as it allows us to separately capture distortions on different scales, to which the human visual system typically has different sensitivity [29]. MS-$\mathrm{C}_{\mathrm{MSE}}^{\mathrm{3D}}$, MS-$\mathrm{C}_{\mathrm{T-MSE}}^{\mathrm{3D}}$, MS-$\mathrm{C}_{\mathrm{DSSIM}}$ and MS-$\mathrm{C}_{\mathrm{MSE}}$ are defined on the basis of (6) and (7), similarly to (3).

Although MS-$\mathrm{C}_{\#}^{\mathrm{3D}}$ can capture motion inconsistencies owing to the use of 3D patches, MS-$\mathrm{C}_{\#}$ detect only spatial errors. To overcome this problem, we propose measuring how the distance to the most similar patch in the ground-truth video changes from frame to frame. More precisely, we find for a given patch the most similar patch from the previous frame within a certain window, compute the distances from these patches to the most similar ground-truth patches and then compare the respective distances. Formally,

$$\begin{aligned}\mathrm{C}_{\#\mathrm{d}t}(V, V_{\mathrm{ref}}) &= \frac{1}{|\Omega|}\sum_{x\in\Omega}\Big|\min_y\Big(\#\left(P(x), P_{\mathrm{ref}}(y)\right)\Big) - \min_y\Big(\#\left(P(x_{\mathrm{prev}}), P_{\mathrm{ref}}(y)\right)\Big)\Big|,\\ x_{\mathrm{prev}} &= \mathrm{argmin}_{y\in\Omega_{\mathrm{prev}}^{w\times w}(x)}\#(P(x), P(y)), \#\\ &= \mathrm{DSSIM}, \mathrm{MSE}\end{aligned} \tag{9}$$

$\Omega_{\mathrm{prev}}^{w\times w}(x)$ is a square window of $w\times w$ pixels (we use w equal to 1/10th of the frame width) spatially centered at x and located in the previous frame. Multiscale variants MS-$\mathrm{C}_{\mathrm{MSEdt}}$ and MS-$\mathrm{C}_{\mathrm{DSSIMdt}}$ are defined similarly to (5). These metrics essentially capture the changes in patch appearance from frame to frame, as opposed to evaluating consistency with ground-truth optical flow using MS-MSEdt and MS-DSSIMdt (5).

Exact computation of coherence-based metrics quickly becomes impractical for larger spatiotemporal holes, so we resort to approximate solutions based on PatchMatch [2]. For metrics that use 3D patches, we generally follow [20]. In the case of metrics that search for the most similar 2D patch, we apply PatchMatch frame by frame—that is, we perform several propagation and random-search iterations to find the nearest-neighbor field (NNF) between the current inpainted frame and the whole ground-truth video before pro-

ceeding to the next frame. At that point, the previous frame result serves as the NNF initialization. In our experiments, PatchMatch converged fairly quickly; we used four spatial-propagation and random-search iterations for all metrics and test sequences. Also, we always use a sum of squared differences (SSD) in the RGB color space when computing NNFs. Explicit optimization of the SSIM metric tends not to be worth the greater computation time.

3.3 Assessing metric performance

Very few works in this field have gone so far as to publish source code, so few video-completion methods are available for evaluation and for assessing metric performance. To increase the amount of data, we included several commercial tools and image-inpainting algorithms. In particular, we evaluated six methods:

- Video Inpainting of Complex Scenes [20]
- Nuke F_RigRemoval tool [26]
- YUVsoft Background Reconstruction tool (BGR) [33]
- PFClean Remove Rig tool [21]
- A simple image-inpainting method by Telea [25]
- An image-inpainting method by Huang et al. [13]

To extend the data set even further we added synthetic results obtained by direct joint optimization of MS-C_{MSE} and MS-C_{MSEdt}. The motivation for this approach is twofold. First, as opposed to metrics based on 3D patches (6), none of the tested methods explicitly optimizes the 2D-patch-based measures on which MS-C_{MSE} and MS-C_{MSEdt} are based. Second, we wanted to verify that lower MS-C_{MSE} and MS-C_{MSEdt} values correspond to higher perceptual quality. This relationship isn't obvious, as these metrics do not explicitly take into account any motion similarity. For optimization, we employed a multiscale reconstruction framework that is very similar to the one in [20], with one important distinction: it employs ground truth for reconstruction. Figure 3 illustrates the procedure. To achieve joint metric optimization, we slightly modify the reconstruction step to employ both the interframe nearest-neighbor field $\mathrm{NNF}_{t\to(t-1)}$ and the nearest-neighbor field $\mathrm{NNF}_{t\to\mathrm{GT}}$ that connects each 2D patch in the current frame of the completion result to the most similar patch from the whole ground-truth video:

$$V[x] = \frac{(1-\tau)\sum_{y\in P(x)} w^s(y) V_{\mathrm{ref}}\left[x+\mathrm{NNF}_{t\to\mathrm{GT}}[y]\right] + \tau \sum_{y\in P(x)} w^t(y) V\left[x+\mathrm{NNF}_{t\to(t-1)}[y]\right]}{(1-\tau)\sum_{y\in P(x)} w^s(y) + \tau \sum_{y\in P(x)} w^t(y)} \tag{10}$$

Here x denotes the currently reconstructed pixel, and τ is the weight of the temporal component, which employs the previously reconstructed frame to reconstruct the current one (we use $\tau = 0.4$). The functions $w^s(y)$ and $w^t(y)$ give higher weights to NNF offsets with lower patch distance. We use an exponential weighting function, as in [20]. This approach enables us to find a local optimum that can depend on the initial guess. We used results from six methods as the initial guesses, achieving in all cases a substantial decrease in both metric values (see Fig. 4). Inclusion of these synthetic results doubles the size of our data set.

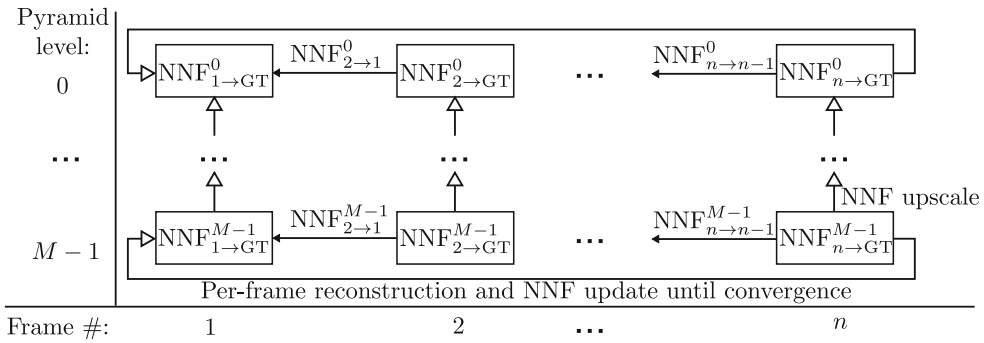

Fig. 3 Block diagram of a simple greedy algorithm that we use to jointly optimize the MS-C_{MSE} and MS-C_{MSEdt} metrics, given the initial completion guess. We iteratively compute interframe nearest-neighbor fields (NNFs) as well as NNFs from the current frame to the whole ground-truth video and then use them for per-frame reconstruction. We compute NNFs using the PatchMatch algorithm [2]. After convergence, we upscale the NNFs and perform the same process for the next pyramid level. Section 3.3 provides more details about the reconstruction step

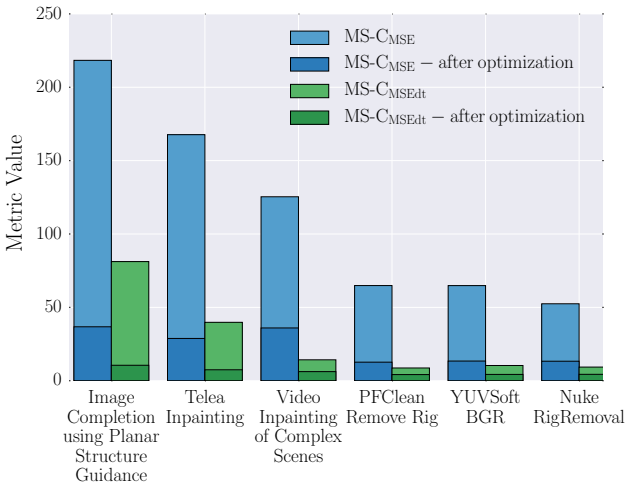

Fig. 4 MS-C$_{MSE}$ and MS-C$_{MSEdt}$ values averaged over all sequences before and after the optimization described in Sect. 3.3. We determined the scale weights as Sect. 4 describes

To evaluate the metrics' performance, we have conducted a study with over 300 paid participants using Subjectify.us web service [24]. Participants were presented with a sequence of video pairs shown side-by-side in a web application; for each pair we asked to select a video with better quality or mark them as equal. For this study, we divided the original dataset of seven sequences into a total of 19 shorter cropped fragments of 70 to 100 frames and 960 × 540 resolution to highlight differences in quality for various parts of the inpainted region. Each of these fragments was used to compare the results of 6 original methods, as well as 6 respective optimized variants and ground truth, resulting in 2964 video pairs that required comparison. In total, we have collected 8,533 pairwise comparisons from 341 participants, distributed uniformly among all methods and test sequences. In particular, each participant viewed 28 video pairs, three of which were hidden verification questions that asked to compare ground-truth result with a very low-quality video completion; only the results of those who correctly answered all three verification questions were used in the study. We offered $0.05 to each participant who had successfully completed the test. Collected pairwise method comparisons were transformed into subjective ranks using Thurstone's Case V Model [27]—both for each test sequence individually and for all sequences at once (Fig. 5). We assess performance of different quality metrics in terms of correlation with these perceptual ranks.

4 Results

To obtain optimal scale weights for multiscale metrics (we use a Gaussian pyramid with five levels), we optimize the correlation with subjective ranks, constrain the sum of the weights to one and impose L_2 regularization:

$$\boldsymbol{w} = \operatorname{argmax}_{\boldsymbol{w}=[w_0,\ldots,w_4] \atop \sum w_i = 1} \sum_{s=0}^{\#\text{sequences}} \operatorname{corr}\big(-\log(\boldsymbol{w}\boldsymbol{M}_s), \boldsymbol{r}_s \big) - \alpha ||\boldsymbol{w}||_2^2 \tag{11}$$

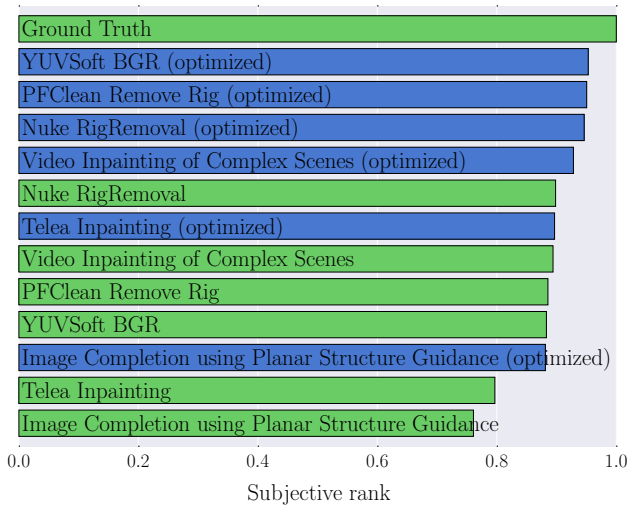

Fig. 5 Overall ranking of different methods computed using pairwise subjective comparisons. We obtained the optimized versions by applying the procedure described in Sect. 3.3, using the results of the respective methods as initial guesses

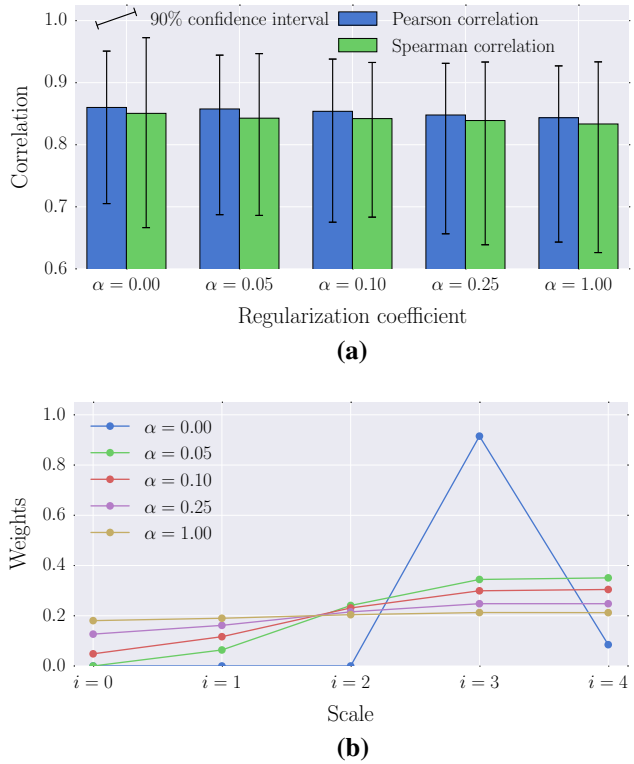

Fig. 6 Illustration of how regularization in (11) affects scale-weight distribution and correlation with subjective ranks. We use $\alpha = 0.1$ in all metrics. **a** Correlation of MS-DSSIM with subjective ranks. **b** Distribution of weights between scales in MS-DSSIM metric

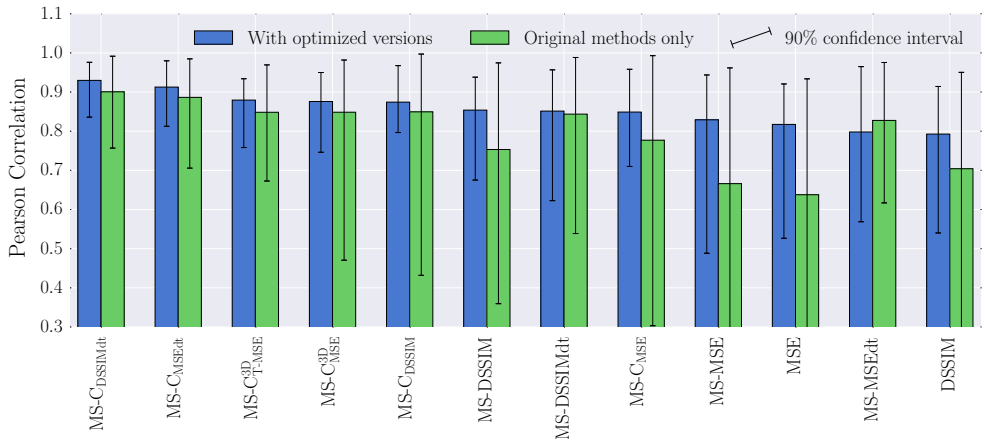

Fig. 7 Correlation of resulting metrics and subjective ranks computed for two different data sets: a full one, including synthetic results obtained by metric optimization using each method's result as the initial guess, and a partial one that includes only the results of six original methods

Here $\boldsymbol{M}_s = \{m^s_{ij}\}$, where m^s_{ij} is a metric value for the ith scale, jth algorithm and sth sequence; $\boldsymbol{r}_s$ contains the corresponding subjective ranks. Figure 6 illustrates the effect of regularization, which makes the metrics more robust by forcing them to cover more scales at the cost of slightly smaller correlation values. The weights can be selected more accurately by synthesizing examples exhibiting a certain type of distortion on different scales, as MSSSIM does [29]. This approach, however, is too labor intensive for our purposes, so we obtain weights for all metrics by simply optimizing (11) with $\alpha = 0.1$. Figure 7 shows the resulting correlation values. First, note that practically all metrics exhibit worse performance when measuring correlation for the data set of six original methods—especially metrics that directly estimate adherence to ground truth (MSE, DSSIM and their multi-scale variants). This behavior occurs because the proposed optimization procedure brings the results closer to ground truth (which is not that surprising, as it uses the ground-truth results directly) and improves the subjective rankings at the same time (see Fig. 5). As a result, the performance of such straightforward metrics receives a boost on the full data set. Regardless of the data set, we observe that SSIM-based metrics consistently show higher correlation with subjective ranks than do their MSE-based counterparts. Also, temporal-based metrics typically perform better than metrics that simply measure spatial error, which is to be expected. On the other hand, MS-C$_{DSSIMdt}$ and MS-C$_{MSEdt}$ perform surprisingly well, given that they only capture changes in patch appearance over time and don't explicitly assess motion similarity. Metrics based on 3D patches also show high correlation with perceptual data, but they are more computation and memory intensive than approaches based on 2D patches. Therefore, we chose the following metrics for the benchmark:

- MS-C$_{DSSIMdt}$, $[w_i] = [0.00, 0.08, 0.25, 0.30, 0.37]$
- MS-C$_{DSSIM}$, $[w_i] = [0.04, 0.11, 0.21, 0.29, 0.35]$
- MS-DSSIMdt, $[w_i] = [0.00, 0.00, 0.30, 0.32, 0.38]$
- MS-DSSIM, $[w_i] = [0.05, 0.12, 0.23, 0.30, 0.30]$

MS-C$_{DSSIMdt}$ and MS-C$_{DSSIM}$ are significantly more computationally complex than conventional metrics (several seconds per 1920×1080 frame in our experiments), but they provide better correlation with subjective ranks. Metric values for all tested methods and sequences, as well as an extensive set of comparative charts, are available at http://videocompletion.org. We also publish all of our test sequences and the results of the tested methods for visual comparison.

5 Conclusion

We have presented an online video-completion benchmark available at http://videocompletion.org. We introduced four objective quality metrics that outperform naïve approaches according to subjective video-completion-method evaluations involving more than 300 participants. Our proposed benchmark metrics have varying levels of correlation with perceptual data, but each has its own intuitive interpretation. We believe that presented benchmark can help rank existing video-completion methods and assist in developing new approaches.

References

1. Anbarjafari, G.: An objective no-reference measure of illumination assessment. Meas. Sci. Rev. **15**(6), 319–322 (2015)

2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. (TOG) **28**(3), 24 (2009)
3. Benoit, J., Paquette, E.: Localized search for high definition video completion. J. WSCG (2015)
4. Blender. https://www.blender.org/
5. Čadík, M., Herzog, R., Mantiuk, R., Myszkowski, K., Seidel, H.P.: New measurements reveal weaknesses of image quality metrics in evaluating graphics artifacts. ACM Trans. Graph. (TOG) **31**(6), 147 (2012)
6. Chen, Y., Hu, Y., Au, O.C., Li, H., Chen, C.W.: Video error concealment using spatio-temporal boundary matching and partial differential equation. IEEE Trans. Multimed. **10**(1), 2–15 (2008)
7. Cheng, E., Burton, P., Burton, J., Joseski, A., Burnett, I.: RMIT3DV:Pre-announcement of a creative commons uncompressed HD 3D video database. In: Fourth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 212–217 (2012)
8. Ebdelli, M., Meur, O.L., Guillemot, C.: Video inpainting with short-term windows: application to object removal and error concealment. IEEE Trans. Image Process. (TIP) **24**(10), 3034–3047 (2015)
9. Erofeev, M., Gitman, Y., Vatolin, D., Fedorov, A., Wang, J.: Perceptually motivated benchmark for video matting. In: British Machine Vision Conference (BMVC) (2015)
10. Erofeev, M., Vatolin, D.: Automatic logo removal for semitransparent and animated logos. Proceedings of GraphiCon **2011**, 26–30 (2011)
11. Granados, M., Tompkin, J., Kim, K., Grau, O., Kautz, J., Theobalt, C.: How not to be seen–object removal from videos of crowded scenes. Comput. Graph. Forum **31**, 219–228 (2012)
12. He, K., Sun, J.: Statistics of patch offsets for image completion. In: European Conference on Computer Vision (ECCV), pp. 16–29 (2012)
13. Huang, J.B., Kang, S.B., Ahuja, N., Kopf, J.: Image completion using planar structure guidance. ACM Trans. Graph. (TOG) **33**(4), 129 (2014)
14. Ilan, S., Shamir, A.: A survey on data-driven video completion. Comput. Graph. Forum **34**, 60–85 (2015)
15. Koloda, J., Ostergaard, J., Jensen, S.H., Peinado, A.M., Sanchez, V.: Sequential error concealment for video/images by weighted template matching. In: Data Compression Conference (DCC), pp. 159–168 (2012)
16. Koloda, J., Ostergaard, J., Jensen, S.H., Sanchez, V., Peinado, A.M.: Sequential error concealment for video/images by sparse linear prediction. IEEE Trans. Multimed. **15**(4), 957–969 (2013)
17. Mosleh, A., Bouguila, N., Hamza, A.B.: Video completion using bandlet transform. IEEE Trans. Multimed. **14**(6), 1591–1601 (2012)
18. Mosleh, A., Bouguila, N., Hamza, A.B.: Automatic inpainting scheme for video text detection and removal. IEEE Trans. Image Process. (TIP) **22**(11), 4460–4472 (2013)
19. Mosleh, A., Bouguila, N., Hamza, A.B.: Bandlet-based sparsity regularization in video inpainting. J. Vis. Commun. Image Represent. **25**(5), 855–863 (2014)
20. Newson, A., Almansa, A., Fradet, M., Gousseau, Y., Pérez, P.: Video inpainting of complex scenes. SIAM J. Imaging Sci. **7**(4), 1993–2019 (2014)
21. Pixel Farm PFClean. http://www.thepixelfarm.co.uk/pfclean/
22. Seshadrinathan, K., Bovik, A.C.: Motion tuned spatio-temporal quality assessment of natural videos. IEEE Trans. Image Process. (TIP) **19**(2), 335–350 (2010)
23. Shiratori, T., Matsushita, Y., Tang, X., Kang, S.B.: Video completion by motion field transfer. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) **1**, 411–418 (2006)
24. Subjectify.us. http://subjectify.us
25. Telea, A.: An image inpainting technique based on the fast marching method. J. Graph. Tools **9**(1), 23–34 (2004)
26. The Foundry Nuke. https://www.thefoundry.co.uk/products/nuke/
27. Thurstone, L.L.: A law of comparative judgment. Psychol. Rev. **34**(4), 273 (1927)
28. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. (TIP) **13**(4), 600–612 (2004)
29. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers **2**, 1398–1402 (2003)
30. Wexler, Y., Shechtman, E., Irani, M.: Space-time completion of video. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **29**(3), 463–476 (2007)
31. Yan, W.Q., Wang, J., Kankanhalli, M.S.: Automatic video logo detection and removal. Multimed. Syst. **10**(5), 379–391 (2005)
32. You, S., Tan, R.T., Kawakami, R., Ikeuchi, K.: Robust and fast motion estimation for video completion. In: International Conference on Machine Vision Applications (MVA), pp. 181–184 (2013)
33. YUVSoft Background Reconstruction. http://www.yuvsoft.com/stereo-3d-technologies/background-reconstruction/