# Localized Search for High Definition Video Completion

Jocelyn Benoit
École des arts numériques, de l'animation et du design
Montreal, Canada
jbenoit@nad.ca

Eric Paquette
Multimedia Lab, École de technologie supérieure
Montreal, Canada
eric.paquette@etsmtl.ca

## ABSTRACT

This paper presents a new approach for video completion of high-resolution video sequences. Current state-of-the-art exemplar-based methods that use non-parametric patch sampling work well and provide good results for low-resolution video sequences. Unfortunately, because of memory consumption problems and long computation times, these methods handle only relatively low-resolution video sequences. This paper presents a video completion method that can handle much higher resolutions than previous ones. First, to address the problem of long computation times, a dual inpainting-sampling filling-order completion method is proposed. The quality of our results is then significantly improved by a second innovation introducing a coherence-based matches refinement that conducts intelligent and localized searches without relying on approximate searches or compressed data. Finally, with respect to the computation times and memory problems that prevent high-resolution video completion, the third innovation is a new localized search completion approach, which also uses uncompressed data and an exact search. Combined together, these three innovations make it possible to complete high-resolution video sequences, thus leading to a significant increase in resolution as compared to previous works.

## Keywords

Video completion, high-resolution, object removal, patches coherence, localized search, multi-resolution

## 1 INTRODUCTION

Both image and video completion are important tasks in many multimedia applications. Their goal is to automatically fill missing regions of an image/video in a visually plausible manner. Two key factors differentiate video completion from image completion. Firstly, for video completion, it is important to maintain temporal consistency since human vision is more sensitive to temporal artifacts than to spatial artifacts. Using an image completion technique individually on each frame produces undesired temporal artifacts. Secondly, it is more important for video completion to be time- and memory-efficient since video contains much more data than image.

In the past years, many new solutions have been proposed for video completion. It has been shown that exemplar-based methods, that use *non-parametric patch sampling*, work well and provide good results. Unfortunately, they work only on relatively low-resolution videos because larger ones require too much memory. Few methods [8, 12] present results for $640 \times 480$ or $540 \times 432$ resolutions, with most [6, 7, 9–11, 15, 17, 20, 21] presenting results of $320 \times 240$ or lower resolutions. Since High Definition (HD) videos with $1920 \times 1080$ or higher resolutions are now commonplace, most of these methods cannot be applied directly or they require too long computation times.



To understand the proposed method, we must first look at the non-parametric patch sampling approaches. Those methods are based on an iteration through each of the patches in the missing regions and a search in all of the patches of the existing regions to find the most similar patch. Without optimization, this search can be excessively time consuming: $O(m^3M^2F)$ with $M$ representing the video width and height; $m$ the patch width, height and depth; and $F$ the number of frames. Even with optimization methods, the search time of the non-parametric patch sampling approaches still remains excessive. Furthermore, the structures needed for theses optimization methods require too much memory, making them inappropriate for HD videos.

Rather than focusing on the acceleration of the nearest neighbors search, the proposed method narrows the search space at finer (higher) resolutions using information obtained at coarser (lower) resolutions. First, let us consider two patches at coarser resolutions: patch $w_p^l$ from the missing region and its most similar patch $w_{p'}^l$ from the existing region. The most sim-

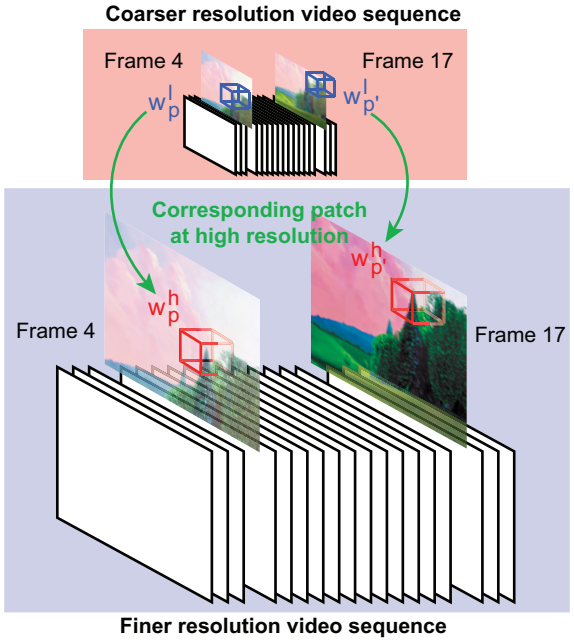


Figure 1: First row: coarser resolution video. Patch ($w_p^l$) in the missing region and its most similar patch ($w_{p'}^l$) in the existing region. Second row: finer resolution video. The corresponding patch ($w_p^h$) of $w_p^l$ and its most similar patch ($w_{p'}^h$) in the existing region.

ilar patch of the corresponding patch $w_p^h$ at finer resolutions is likely to be found near $w_{p'}^h$, as illustrated in Figure 1. The proposed approach begins by completing the video at coarser resolutions using a dual inpainting-sampling filling-order completion approach based on Wexler et al. [20]. Since efficient but approximate search approaches are used to find the most similar patches, errors are introduced and several matches are sub-optimal patches. To solve this problem, a coherence-based matches refinement process is used to search for better matches. The technique then stores the space-time location of the most similar patch found for each patch of the missing region in a matches list *ML*. This *ML* is then used by a localized search completion approach to narrow the search space in higher resolution, thus enabling the completion of HD video sequences.

The contributions of the proposed method include a dual inpainting-sampling filling-order completion approach based on Wexler et al. [20]; a new coherence-based matches refinement process that improves the quality of the matches when approximate search approaches are used; and a new localized search completion approach based on an exact search using uncompressed data but restricted to a localized region. We show that the proposed methods enable the completion of HD video sequences and that they produce visually plausible results within reasonable timeframes. Moreover, the approach requires very little memory at the finest resolution except for the input video storage.

## 2 PREVIOUS WORKS

In past years, many methods have been proposed to replace missing regions of an image. *Image inpainting* techniques propose to fill the missing region by extending the surrounding existing region until the hole vanishes. These techniques generally work only on small and thin holes. *Image completion* techniques use non-parametric patch sampling and are able to fill even larger missing regions of an image. While video completion methods are based on image completion and inpainting methods, video completion poses the additional challenge of maintaining spatio-temporal consistency. Using image completion or image inpainting methods on each frame independently produces temporal artifacts that are easily noticed by the viewer [3].

### 2.1 Video completion

Extending the image completion methods based on Markov Random Fields (MRF) and non-parametric patch sampling, Wexler et al. [19, 20] address the problem of video completion as a global optimization, and thus obtain good results on relatively large missing regions. Shiratori et al. [17] proposed a similar approach, but find patches based on motion fields instead of color values. Xiao et al. [21] extend these works by formulating video completion as a new global optimization problem defined over a 3D graph defined in the space-time volume of the video. Liu et al. [10] later have proposed an algorithm with two stages: motion fields completion and color completion via global optimization. The major drawback of all these approaches is the amount of information that must be processed when considering HD video sequences. While some methods use per-pixel searches [19–21], other approaches use larger primitives instead of pixels: Shih et al. [16] use fragments, while Cheung et al. [4] use “epitomes”. Approaches using fragments or epitomes can reduce the search time and improve overall coherence, but per-pixel searches are more likely to correctly restore the fine and subtle details found in HD video sequences.

Many methods segment the video sequence into foreground and background parts [6–8, 11, 14] or into layers [22]. These methods create a static background mosaic of the entire sequence, and as a result, these techniques are limited to video sequences with a static background using a fixed camera. Patwardhan et al. [12] later proposed a framework for dealing with videos containing simple camera motions, such as small parallax and hand-held camera motions. The major drawback with all these techniques is that the pixels replaced are static across the video sequence, thus removing details such as video noise, film grain, or slightly moving objects, such as tree leaves, from the background. At an

HD resolution, this lack of detail is quickly noticed by the viewer.

## 2.2 Coherence techniques

When Ashikhmin [2] introduced the concept of coherence, he observed that the results of synthesis algorithms often contain large contiguous regions of the input texture/image when using non-parametric patch sampling. Consequently, the independent search for every patch in the input texture/image can be accelerated by using information from previously computed searches. Thus it limits the search space of a given patch to the locations of the most similar patches of its neighbors. We based our coherence-based matches refinement on the same coherence observation and developed an novel approach that is efficient with respect to both computation time and memory consumption. Tong et al. [18] also proposed a coherence technique called *k-coherence*. While this technique improved the search time, the pre-processing time and the memory consumption are major drawbacks for high-resolution video completion methods.

This paper presents an approach for the completion of video sequences that requires very low memory usage and reasonable computation time, making it usable for HD video sequences. Further, it presents a new coherence-based match refinement approach that increases the overall quality of the results by eliminating many noticeable artifacts. Unlike most of the previous works, this paper presents results on video sequences with non-stationary camera movements.

# 3 HIGH DEFINITION VIDEO COMPLETION

In this section, we present a new video completion approach that is able to automatically fill missing regions of HD video sequences. Section 3.1 presents the approach overview, Section 3.2 explains the dual inpainting-sampling filling-order completion approach, Section 3.3 describes the coherence-based matches refinement process, and Section 3.4 details the new localized search completion method.

## 3.1 Approach overview

Starting with an input video sequence $V$ containing a missing region or hole $H$ ($H \subset V$), our approach fills $H$ in a visually plausible manner by copying similar patches found in the existing region $E$ ($E = V \setminus H$), thus creating a completed video sequence $V^*$. This process is shown in Figure 2. In order to maintain spatio-temporal consistency, we consider the input video as a space-time volume, and thus a pixel located at $(x, y)$ in frame $t$ can be represented by the space-time point $p = (x, y, t)$. Consenquently, a patch $w_p$ can be seen as

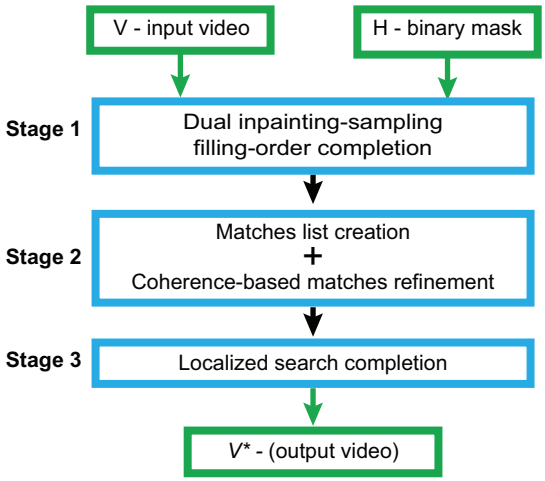


Figure 2: Schematic overview of the proposed approach

| | |
|---|---|
| $V$ | original video sequence |
| $H$ | missing region or hole of $V$, $H \subset V$ |
| $E$ | existing region of $V$, $E = V \setminus H$ |
| $H^*$ | completed region |
| $V^*$ | completed video sequence |
| $p_l$ | point located at (x, y, t) at coarser resolution |
| $w_p^l$ | patch centered at $p_l$ at coarser resolution |
| $w_{p'}^l$ | patch centered at $p'_l$, most similar to patch $w_p^l$ |
| $p_h$ | point located at (x, y, t) at finer resolution |
| $w_p^h$ | patch centered at $p_h$ at finer resolution |
| $w_{p'}^h$ | patch centered at $p'_h$, most similar to patch $w_p^h$ |
| $ML$ | matches list |
| $c$ | RGB color |
| $S$ | search region centered at $p'_h$ |

Table 1: Symbols definitions

a spatio-temporal cube of pixels centered at $p$. Table 1 summarizes the symbols used in this paper.

The missing region $H$ is indicated to the system by a binary video sequence in which identified pixels are in $H$. The binary video sequence can be constructed using object tracking in the video sequence. Many digital motion graphics and compositing softwares already provide accurate and rapid tools to create such binary video sequences. In our experimentations, we relied on such tools to define $H$.

Figure 2 shows a schematic overview of our approach while Figure 3 presents the detailed steps. First, the input video is downsampled and completed by a dual inpainting-sampling filling-order completion based on the works of Wexler et al. [19, 20] using global optimization and non-parametric patch sampling (see Section 3.2). Completing the video at low-resolution with the proposed approach is efficient and provides good results. When the dual inpainting-sampling filling-order completion is finished, each patch $w_p^l$ (centered at $p_l \in H$) is associated with its best matching patch $w_{p'}^l$ (centered at $p'_l \in E$) by creating a matches list containing

space-time pairs $p_l$-$p'_l$ for every $p_l \in H$. To complete the search in a reasonable amount of time, the best matching patches $w^l_{p'}$ must be selected using approximate search and data compression methods. Thus, the $w^l_{p'}$ found might not be the best match. The second stage of the proposed method consists of an iterative coherence-based matches refinement process that improves the search results for the worst matching patches $w^l_{p'}$ (see Section 3.3). This stage is efficient, and provides significant quality improvement. Finally, the matches list is used by the localized search completion method to narrow the search space, thus enabling the completion of HD video sequences (see Section 3.4). This final stage of the method is also efficient, and it provides good results at HD resolution.

## 3.2 Dual inpainting-sampling filling-order completion

A visually plausible completion of a video sequence replaces the missing region $H$ by a completed region $H^*$ where pixels of $H^*$ fit well within the whole video $V^*$. To achieve this, a video completion approach must satisfy two criteria: first, every local space-time patch of the completed region $H^*$ must be similar to an existing patch of $E$, and secondly, all patches that fill $H^*$ must be coherent with each other. Consequently, we seek a completed video $V^*$ that minimizes the objective function stated in Equation 1:

$$Coherence(H^*|E) = \prod_{p_l \in H} \min_{p'_l \in E} D(w^l_p, w^l_{p'}), \quad (1)$$

where $D(w^l_p, w^l_{p'})$ is a similarity metric between two patches. The similarity value of two patches is evaluated with the Sum of Squared Differences (SSD) of color information (in the RGB color space) for every pair of space-time points contained in these patches. Wexler et al. [20] added the spatial and temporal derivatives to the RGB components to obtain a five-dimensional representation for each space-time point. In experimentations, RGB alone produced good results for most videos. Problems occured when trying to reconstruct a hidden moving object. While the technique of Wexler et al. [20] can solve these problems, it is however limited to objects with cyclic motion (i.e. like a walking person). Moreover, it requires more memory and computation time. For these reasons, we limited our problem domain to videos without occluded moving objects and chosen to use only RGB components.

The first step of the dual inpainting-sampling filling-order completion approach is to downsample $V$ to a coarse resolution (see Figure 3, Stage 1.1). Then, before starting the completion, the values of each space-time point of $H$ need to be initialized. Unlike Wexler et al. [20] who used random values, the proposed approach fills $H$ using an image inpainting technique [3]

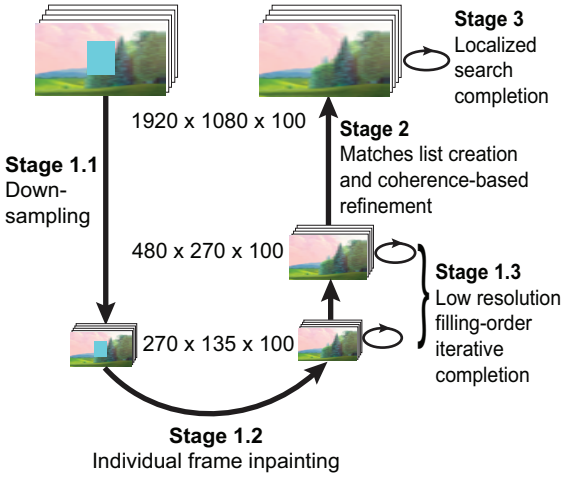


Figure 3: Steps of the proposed video completion approach

(see Figure 3, Stage 1.2). Our aim is to speed up the convergence by using the existing information around $H$. This initialization is done only once, prior to the first iteration of the low resolution filling-order iterative completion approach.

After the initialization, the approach performs an iterative process, improving the overall coherence of $H$ (see Figure 3, Stage 1.3). During each iteration, the approach seeks a replacement color value for every space-time point in $H$ in order to minimize Equation 1. Unlike previous methods, which used scan-line ordering, our approach fills $H$ using a 3D hole-filling approach, thus ensuring that each patch $w^l_p$ contains information that is more reliable (space-time points in $E$ or space-time points already processed during the current iteration). Consequently, it speeds-up the convergence and reduces discontinuities near the boundaries of $H$. The patches can have different sizes in the spatial and temporal dimensions. Generally, we used $5 \times 5 \times 5$ patches or $7 \times 7 \times 5$ patches and we based our choice on the element structure size that needs to be completed within the video sequence.

To seek a replacement color $c$ for a space-time point $p$, the approach uses a single best-matching patch $w^l_{p'}$ that minimizes $D(w^l_p, w^l_{p'})$. When $w^l_{p'}$ is found, the color $c'$ is copied from space-time point $p'_l$ to $p_l$. Compared to other methods that blend together several matches, using the single best-matching patch does not result in blurring artifacts and preserves film grain and noise from the original video. For these reasons, our approach uses the single best-matching patch.

To enforce spatio-temporal consistency, this iterative process is done on multiple scales using spatial pyramids (see Figure 3, Stage 1.3). Each pyramid level contains $^1/_2 \times {}^1/_2$ of the spatial resolution while maintaining the temporal resolution. The iterative process starts with the coarsest pyramid level and propagates its re-


Figure 4: Comparison of the results obtained with the low-resolution video completion approach: (a) Original frames; (b) results from Wexler et al. [20]; (c) results from the proposed method

sults to finer levels. Because it involves long computation times and a lot of memory for the search structure, this iterative process is impractical at finer pyramid levels for HD videos. Therefore, the proposed approach stops the iterative process when it reaches a fixed resolution (typically $480 \times 270$).

The proposed dual inpainting-sampling filling-order completion approach produces results with a quality equivalent to the results of Wexler et al. [20], but within much less time. Figure 4 shows the completion results of the "Jogging lady" sequence of Wexler et al. [20] and ours. Wexler's approach took one hour per iteration at the finest resolution level while our approach took less than four minutes per iteration.

## 3.3 Coherence-based matches refinement

When Stage 1 is over, each patch $w_p^l \in H^*$ has a corresponding patch $w_{p'}^l$. Each space-time point $p_l$ is associated with its corresponding $p'_l$ and the pairs are stored in a matches list *ML*. During the high-resolution completion iterative process, *ML* enables the approach to narrow the search space to only sub-regions of *E*. As a reminder, our key observation is that, for a patch $w_p^l$ at coarser resolution with its most similar patch $w_{p'}^l$, the most similar patch of the corresponding patch $w_p^h$ at finer resolution is likely to be found near $p'_h$ (see Figure 1).

For efficiency reasons, optimization methods such as principal component analysis (PCA) and approximate nearest neighbors search (ANN) [1] are used in Stage 1. While these methods are essential to achieving acceptable search times, they often provide matches $w_{p'}^l$ that do not minimize Equation 1. Consequently, *ML* needs to be refined during Stage 2 (see Figure 3, Stage 2) to have better matching patches $w_{p'}^l$.


Figure 5: Impact of the *ML* refinement iterative process on high-resolution video completion results: (a) Original frame; (b) completed frame without *ML* refinement; (c) completed frame with *ML* refinement. The frames were cropped to better show the missing and completed regions

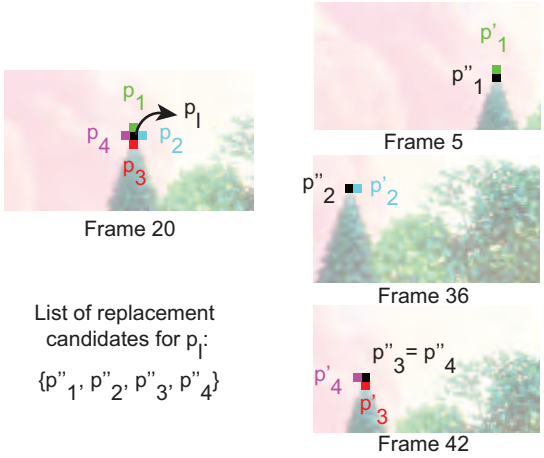


Figure 6: Coherence-based matches refinement process

The information contained in *ML* must be reliable in order for visually plausible results to be possible with the localized search completion iterative process. The patches $w_p^l$ and $w_{p'}^l$ must be highly similar for every pair $p_l$-$p'_l$ of *ML*, otherwise, the approach will narrow the search space to a region where it is less likely to find the best matching $w_{p'}^h$. Figure 5 (a, b) shows an example where the information contained in *ML* is not reliable. As can be seen, there are many visible artefacts such as the centered window and the left building edge.

To find better matching patches $w_{p'}^l$, we take advantage of the concept of coherence. First, the approach calculates the distance ($L_2$ norm of uncompressed data) of patches $w_p^l$ and $w_{p'}^l$ for each pair $p_l$-$p'_l$ from *ML*. Then an iterative process refines pairs with distances higher than a given threshold. During the first iteration, this threshold is set such that 15% of the pairs are refined. After each iteration, this threshold is reduced by 20% of its initial value. For each pair $p_l$-$p'_l$ above the threshold, the approach seeks for a replacement $p'_l$ that decreases $D(w_p^l, w_{p'}^l)$. Instead of using a brute force approach that searches the entire video sequence, the search is restricted around the best matching patches of

$p_l$ neighbors. An example is shown in Figure 6. The four neighbors of $p_l$ are considered: top, right, bottom, and left; respectively $p_1$, $p_2$, $p_3$, and $p_4$. For the top neighbor ($p_1$) the approach considers its previously calculated best matching point $p'_1$, then from $p'_1$, its bottom neighbor $p''_1$ is considered. The $L_2$ norm is computed between patches $p_l$ and $p''_1$, and if the norm is lower than the current value, $p'_l$ is replaced by $p''_1$ and the color from $p''_1$ is copied to $p_l$. This process is repeated for $p_2$, $p_3$, and $p_4$. If there is no good replacement, the pair $p_l$-$p'_l$ is left unchanged, and is considered in the next iteration.

When considering the top neighbor $p_1$, instead of searching anywhere around its best matching point $p'_1$, only its bottom neighbor ($p''_1$) is considered. The rationale behind this is that several successful approaches use large primitives such as fragments or epitomes. When considering larger primitives, the bottom neighbor ($p''_1$) is the one that would be copied on top of $p_l$. This effectively reduces the search to only four points ($p''_1$ to $p''_4$). To even further reduce the number of points to test, each neighbor $p_1$ to $p_4$ is considered only if the $L_2$ norm of a pair, for example, $p_1$-$p'_1$, is below the current threshold. This is a very rapid test since the value is already computed and stored in the *ML*. Since there is a maximum of only four potential points to consider as opposed to the millions from the whole video sequence, this process is extremely fast. Figure 7 shows an example of the *ML* coherence-based matches refinement process that minimizes the distance of $w_p$ and $w_{p'}$ for each pair $p_l$-$p'_l$ in *ML*. The *ML* refinement provides a significant quality improvement (as shown in Figure 5) within a few seconds.

## 3.4 Localized search completion

This section presents the proposed approach for completing missing regions of video sequences at HD resolutions. As stated earlier, current exemplar-based methods are unpratical to complete HD video sequences because best match searches require excessive amount of memory and computation time. Many attempts have been made to accelerate this search with optimization methods such as ANN and dimensionality reduction methods such as PCA, but the structures needed for these optimization methods require too much memory for HD video sequences. Instead of accelerating the best match search, the proposed method narrows the search space at HD resolution using information from coarser resolutions.

Before the localized search completion process starts, the information contained in *ML* must to be scaled up to the finer resolution (see Figure 8). For each space-time point $p_h \in H$ at a finer resolution, its corresponding low-resolution $p_l$ is found as well as the space-time location $p'_l$ associated with it. The space-time location

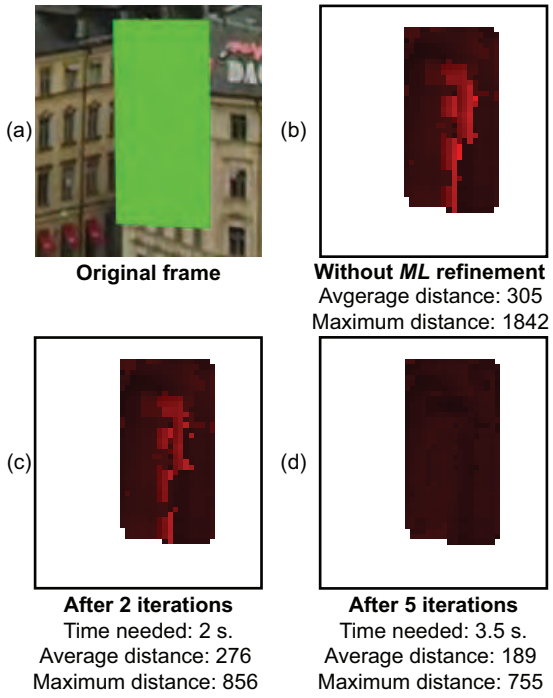


Figure 7: Impact of the *ML* coherence-based matches refinement process: (a) original frame; (b) distance of $w_p$ and $w_{p'}$ for each pair $p_l$-$p'_l$ after *ML* creation; (c) distances after two iterations of the *ML* refinement process; (d) distances after four iterations of the *ML* refinement process. The frames were cropped to better show the missing and completed regions

$p'_l$ is then scaled up to a finer resolution resulting in $p'_h$. The pair $p_h$-$p'_h$ is then added in a new matches list *MLH* which will be used by the localized search completion process to narrow the search space.

The main steps of the localized search completion process are similar to those of the low-resolution process: using a 3D hole-filling approach, the method seeks a

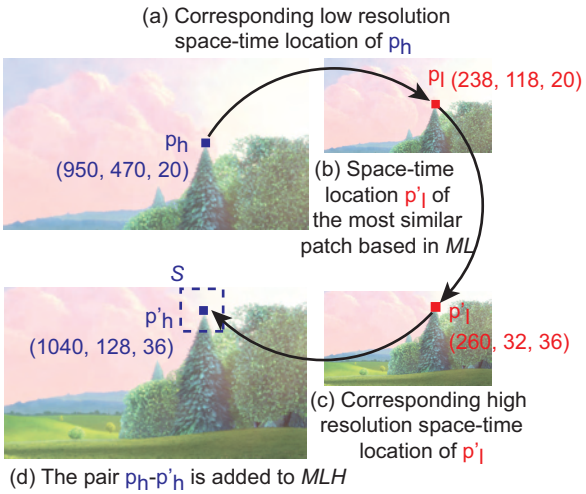


Figure 8: Creation of *MLH* based on *ML*

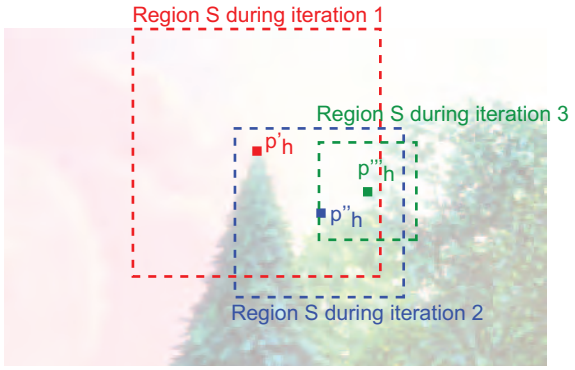


Figure 9: Locations and sizes of search regions $S$ for the first three iterations of the localized search completion process

replacement color $c'$ for every $p_h \in H$ using a single best-matching patch $w^h_{p'}$. However, instead of searching through the entire video sequence using a brute force algorithm or expensive search structures, the approach only searches in a small sub-region $S$, based on the information from *MLH*. For each space-time point $p_h \in H$, the approach first looks in *MLH* and seeks for its associated $p'_h$. Then, a small region $S$ centered at $p'_h$ is selected. Next, the approach searches only in $S$ for the best-matching patch $w^h_{p''}$ (located at $p''_h \in S$) and the color $c_h$ is replaced by $c''_h$. If $p'_h$ and $p''_h$ are different, the pair $p_h$-$p'_h$ from *MLH* is replaced with the pair $p_h$-$p''_h$. In the next iteration of the localized search completion process, the sub-region $S$ will be recentered around this updated space-time location. During the first iteration of the localized search completion process, the window size of sub-region $S$ is $17 \times 17$ pixels. This window size is then decreased after each iteration ($13 \times 13$, $9 \times 9$, $5 \times 5$). Figure 9 shows an example of the location changes and size decreases of a search region $S$ for three iterations.

Obviously, the search time is dramatically reduced when using *MLH* to narrow the search space, as compared to using methods such as ANN and PCA. When using the proposed *MLH* technique, less than a thousand patches are searched for each $p_h$ compared to the tens of millions of patches from the whole video. Moreover, the computation time for the creation and the refinement of *ML* and *MLH* is shorter than the time needed for the creation of the structures used by ANN and PCA. Another important advantage of the proposed method is that *MLH* requires much less memory than typical search structures, such as ANN. Finally, the proposed *MLH* search does not rely on compressed data, and thus can provide better matches.

## 4 RESULTS AND DISCUSSION

Figure 10 shows the completion of the "Station" sequence and Figure 11 shows the completion of the "Race to Mars" sequence. The main challenge of these sequences is the constant motion of the camera. The "Station" sequence contains a constant zooming motion while the "Race to Mars" sequence contains complex rotating and panning motions. Video sequences with such motions cannot be handled by video completion techniques using a static background mosaic because the size and orientation of the objects contained in the background are not constant during the entire video sequence.

It can be seen in Figure 10 that the proposed method works well with large missing regions. Figures 10 and 11 demonstrate that the proposed methods produce good results for missing regions containing stochastic texture as well as salient structure. Since state of the art papers introduced in Section 2 show results with resolutions ranging from $320 \times 240$ to $640 \times 480$, it is not possible to compare the quality of our results with other techniques. Therefore, we used a structural similarity method (SSIM) [13], a full reference metric, to measure the quality of our results at high-resolution. Even though SSIM is generally used to evaluate video compression methods, it can also be used to measure the similarity between a reference sequence and a completed sequence. Figure 12 shows the completion of the "Old town cross" sequence. Considering only the pixels in the missing region instead of all the pixels from the full frames of the sequence, the average SSIM index is 90.63. Since the completed region does not need to be exactly like the reference region, as long as the region is completed in a visually plausible manner, this SSIM index is good.

Table 2 shows a comparison of the proposed approach with earlier works based on different criteria (some were taken from Shih et al. [15]):

- **Missing region specification**: how the user interacts with the method to specify the missing region;
- **Exemplar-based approach**: what type of completion method is used;
- **Camera motion**: video sequences with stationary or nonstationary camera;
- **Maximum resolution**: the highest resolution of the video sequences presented in the paper.

All completion methods use an exemplar-based technique with different variations. Most of the completion methods only use video sequences taken with a stationary camera to test their algorithm. Patwardhan et al. [12] present results with a non-static camera, but the camera motion is always parallel to the projection plane. Thus, Patwardhan et al. [12] do not deal with changes in size, perspective, nor zooming. Only Shih et al. [15] and the proposed method present results with

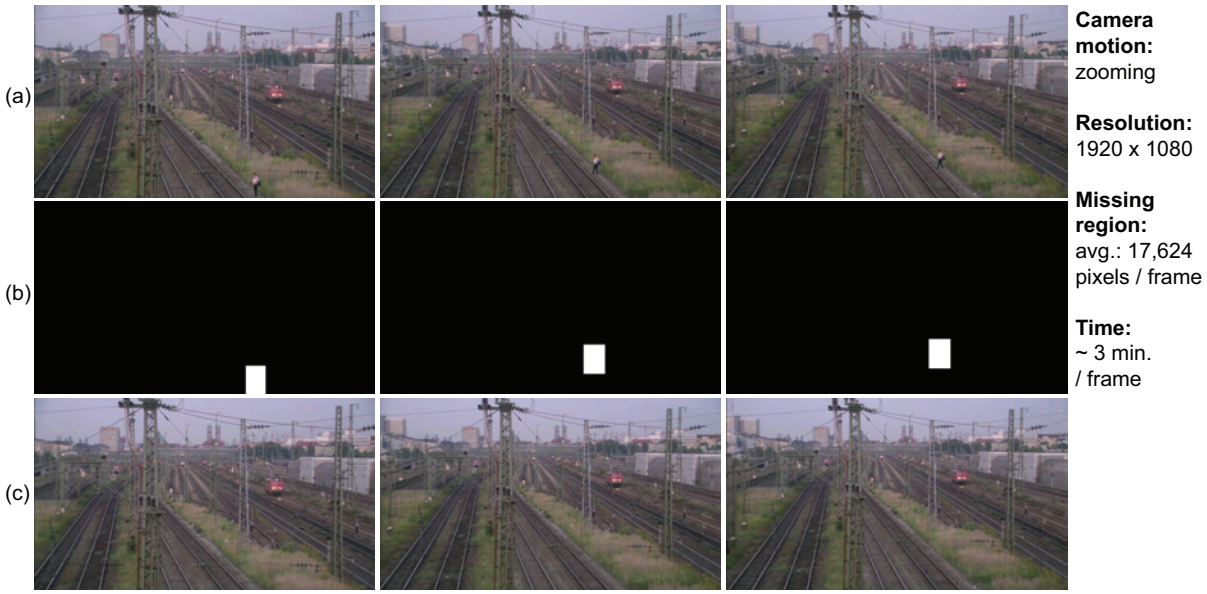


Figure 10: Results for the "Station" sequence : (a) Original frames; (b) missing regions; and (c) completed frames. Frames from `https://cs-nsl-wiki.cs.surrey.sfu.ca/wiki/Video_Library_and_Tools`

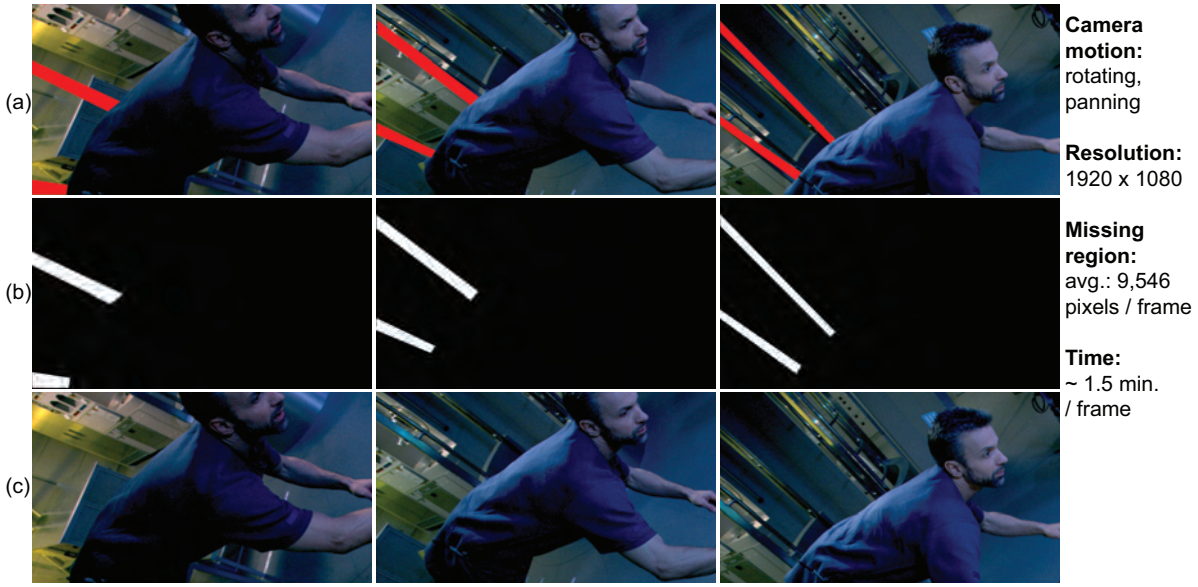


Figure 11: Results for the "Race to Mars" sequence (frames were cropped to better see the regions): (a) Original frames (with unwanted wires highlighted in red); (b) missing regions; and (c) completed frames. Frames from "Race to Mars", a courtesy of Galafilm and Discovery Channel Canada

different camera motions such as zooming, rotating, and panning. Finally, the main advantage of the proposed method over previous works is the maximum resolution it can handle. The proposed method handles HD video sequences while the highest resolution of all previous works from Section 2 is only $640 \times 480$, which is more than a six-fold improvement over state-of-the-art exemplar-based methods.

## 5 CONCLUSION

We have presented a video completion method that can handle much higher resolutions than previous work. The proposed method is based on three new approaches: a dual inpainting-sampling filling-order completion, a new coherence-based matches refinement, and a new localized search completion approach. Together, these three approaches solve the memory consumption and computation time problems for the completion of HD video sequences. Furthermore, the quality of the results generated by our method compares fa-

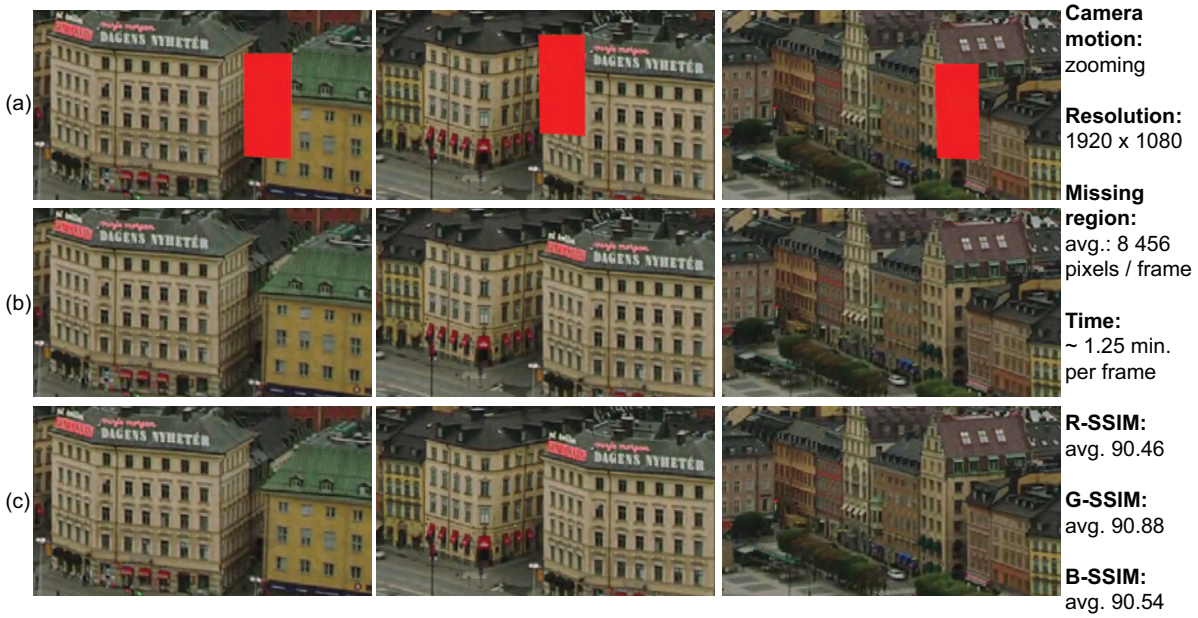


Figure 12: Results for the "Old town cross" sequence (frames were cropped to better see the regions): (a) Frames with a synthetic object; (b) completed frames; and (c) clean frames. Frames from `https://cs-nsl-wiki.cs.surrey.sfu.ca/wiki/Video_Library_and_Tools`

| | Criteria | | | |
|---|---|---|---|---|
| **Related works** | **Missing region(s) specification** | **Exemplar-based approach** | **Camera motions** | **Maximum resolution** |
| Kamel et al. [7] | User provided mask | Standard | Static | $80 \times 110$ |
| Shih et al. [15] | Bounding box given by user, missing region is tracked | Improved patch-matching | Static, non-static | $320 \times 240$ |
| Liu et al. [10] | User provided mask | Motion fields and colors | Static | $320 \times 240$ |
| Xiao et al. [21] | User provided mask | Motion similarity and colors | Static | $384 \times 192$ |
| Shiratori et al. [17] | User provided mask | Motion fields | Static | $352 \times 240$ |
| Wexler et al. [20] | User provided mask | Motion similarity and colors | Static | $360 \times 288$ |
| Koochari and Soryani [8] | User provided mask | Standard | Static | $540 \times 432$ |
| Patwardhan et al. [12] | User provided mask | Motion inpainting and priority based texture synthesis | Static, non-static | $640 \times 480$ |
| Herling and Broll [5] | Rough selection by user, missing region is tracked | Combined pixel-based approach | Static, non-static | $640 \times 480$ |
| The proposed approach | User provided mask | dual inpainting-sampling filling-order completion, coherent and localized search | Static, non-static | $1920 \times 1080$ |

Table 2: Comparison of the proposed method with previous works

vorably to previous works and allows for a significant increase of the resolutions that can be completed.

The proposed coherence-based match refinement is promising as it could be applied at various steps of several video completion approachs. Future work will involve an investigation of when the coherence approach provides the best improvements: between each iterations; between each resolution levels; at coarser or finer resolutions; etc. As they are used in the proposed method, the coherence-based matches refinement and localized search completion consider a fairly limited number of patches. Therefore, the search could stop in a local minimum while there are better matches elsewhere in the video. Future work should look at appropriate techniques to expand the search to other locations that are likely to contain good matches.

## 6 ACKNOWLEDGMENTS

We would like to thank the Fonds québécois de la recherche sur la nature et les technologies (FQRNT),

Mokko Studio inc., and the École de technologie supérieure for funding this project. We would also like to thank all members of the Multimedia Lab for their reviews. Finally, we would also want to thank Galafilm and Discovery Channel Canada for the "Race to Mars" video sequence.

## REFERENCES

[1] Arya, S., Mount, D.M.: A library for approximate nearest neighbor searching (2010). URL `http://www.cs.umd.edu/~mount/ANN/`

[2] Ashikhmin, M.: Synthesizing natural textures. In: 2001 ACM symposium on Interactive 3D graphics, pp. 217–226. ACM (2001)

[3] Bertalmio, M., Bertozzi, A., Sapiro, G.: Navier-stokes, fluid dynamics, and image and video inpainting. In: Proc. Conf. Comp. Vision Pattern Rec., pp. 355–362 (2001)

[4] Cheung, V., Frey, B.J., Jojic, N.: Video epitomes. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 42–49 (2005)

[5] Herling, J., Broll, W.: High-quality real-time video inpaintingwith pixmix. IEEE Transactions on Visualization and Computer Graphics **20**(6), 866–879 (2014)

[6] Jia, J., Tai, Y.W., Wu, T.P., Tang, C.K.: Video repairing under variable illumination using cyclic motions. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(5), 832 –839 (2006)

[7] Kamel, S., Ebrahimnezhad, H., Ebrahimi, A.: Moving object removal in video sequence and background restoration using kalman filter. In: Int. Symp. on Telecommunications 08, pp. 580–585 (2008)

[8] Koochari, A., Soryani, M.: Exemplar-based video inpainting with large patches. Journal of Zhejiang University - Science C **11**(4), 270–277 (2010)

[9] Ling, C., Lin, C., Su, C., Liao, H.Y., Chen, Y.: Video object inpainting using posture mapping. In: Proc. IEEE Int. Conf. on Image Processing, pp. 2785–2788 (2009)

[10] Liu, M., Chen, S., Liu, J., Tang, X.: Video completion via motion guided spatial-temporal global optimization. In: Proc. ACM Multimedia, pp. 537–540 (2009)

[11] Patwardhan, K., Sapiro, G., Bertalmio, M.: Video inpainting of occluding and occluded objects. In: Proc. IEEE Int. Conf. on Image Processing, vol. 2, pp. 69–72 (2005)

[12] Patwardhan, K.A., Sapiro, G., Bertalmio, M.: Video inpainting under constrained camera motion. IEEE Transactions on Image Processing **16**(2), 545 –553 (2007)

[13] Sheikh, H., Sabir, M., Bovik, A.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing, **15**(11), 3440 –3451 (2006)

[14] Shen, Y., Lu, F., Cao, X., Foroosh, H.: Video completion for perspective camera under constrained motion. In: Proc. of the 18th Int. Conf. on Pattern Recognition, vol. 3, pp. 63–66 (2006)

[15] Shih, T., Tang, N., Hwang, J.N.: Exemplar-based video inpainting without ghost shadow artifacts by maintaining temporal continuity. IEEE Transactions on Circuits and Systems for Video Technology **19**(3), 347–360 (2009)

[16] Shih, T.K., Tang, N.C., Yeh, W.S., Chen, T.J.: Video inpainting and implant via diversified temporal continuations. In: Proc. of the 14th annual ACM int. conf. on Multimedia, MULTIMEDIA '06, pp. 965–966. ACM (2006)

[17] Shiratori, T., Matsushita, Y., Tang, X., Kang, S.B.: Video completion by motion field transfer. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 411–418 (2006)

[18] Tong, X., Zhang, J., Liu, L., Wang, X., Guo, B., Shum, H.Y.: Synthesis of bidirectional texture functions on arbitrary surfaces. ACM Trans. Graph. **21**, 665–672 (2002)

[19] Wexler, Y., Shechtman, E., Irani, M.: Space-time video completion. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 120–127 (2004)

[20] Wexler, Y., Shechtman, E., Irani, M.: Space-time completion of video. IEEE Trans. on Pattern Analysis and Machine Intelligence, **29**(3), 463–476 (2007)

[21] Xiao, C., Liu, S., Fu, H., Lin, C., Song, C., Huang, Z., He, F., Peng, Q.: Video completion and synthesis. Computer Anim. Virt. Worlds **19**, 341–353 (2008)

[22] Zhang, Y., Xiao, J., Shah, M.: Motion layer based object removal in videos. In: 17th IEEE Workshops on Application of Computer Vision, vol. 1, pp. 516 –521 (2005)