DOI: 10.1111/j.1467-8659.2012.03000.x
EUROGRAPHICS 2012 / P. Cignoni, T. Ertl
(Guest Editors)

Volume 31 (2012), Number 2

How Not to Be Seen – Object Removal from Videos of Crowded Scenes

M. Granados[1] J. Tompkin[2] K. Kim[1] O. Grau[3] J. Kautz[2] C. Theobalt[1]

[1]MPI Informatik [2]University College London [3]BBC R&D

Figure 1: *To remove the foremost person from this video, both the dynamic scene elements and the background behind it need to be restored. In this sample from our* Museum *sequence, the right-hand-side of each frame pair shows the inpainted result.*

Abstract

Removing dynamic objects from videos is an extremely challenging problem that even visual effects professionals often solve with time-consuming manual frame-by-frame editing. We propose a new approach to video completion that can deal with complex scenes containing dynamic background and non-periodical moving objects. We build upon the idea that the spatio-temporal hole left by a removed object can be filled with data available on other regions of the video where the occluded objects were visible. Video completion is performed by solving a large combinatorial problem that searches for an optimal pattern of pixel offsets from occluded to unoccluded regions. Our contribution includes an energy functional that generalizes well over different scenes with stable parameters, and that has the desirable convergence properties for a graph-cut-based optimization. We provide an interface to guide the completion process that both reduces computation time and allows for efficient correction of small errors in the result. We demonstrate that our approach can effectively complete complex, high-resolution occlusions that are greater in difficulty than what existing methods have shown.

Categories and Subject Descriptors (according to ACM CCS): I.4.4 [Image processing and computer vision]: Image restoration— I.4.9 [Image processing and computer vision]: Applications—

Keywords: video restoration, video completion, video inpainting

1. Introduction

Removing unwanted objects or artifacts from videos is a common task in professional video and movie productions. For instance, when filming in public locations, it is often necessary to remove walking people and other objects that accidentally occlude the scene. Objects may also have to be erased from a video sequence due to copyright issues. In other cases, the film crew needs to be in a scene for technical reasons, and needs to be removed in post-processing.

Removing undesired objects from footage requires *completing* the disoccluded region in a perceptually plausible way (see Fig. 1). On images, this is a difficult problem that often requires manual interaction to achieve coherent results [BSFG09]. On videos, this difficulty is exacerbated due to the sensitivity of the human visual system to temporal artifacts [Wan95]. Furthermore, it is common that in a given scene there are multiple moving objects each occluding or being occluded by the *object to be removed*. As a result, video completion is a tremendously difficult task that requires artists to spend many hours of tedious manual work to remove even small objects. Consequently, a semi-automatic completion tool to assist users in this task is highly desirable. However, building such a tool is very challenging: General video completion is an ill-posed problem, as there is no unique solution for the unobserved, occluded regions.

Fortunately, videos contain a high degree of redundancy, with repetitive patterns occurring at different locations, times, and scales [GBI09]. We propose a non-parametric completion method to exploit this redundancy. It works by computing a 3-dimensional discrete offset field, or *shift-volume*, for the missing region. Each occluded pixel holds an offset to a source pixel that could replace the occluded data. This is formulated using a Markov Random Field that penalizes pairwise differences between the neighborhoods of the sources for the occluded pixels (Sec. 3). Our algorithm builds on the concepts of *correspondence-maps* [DSC03], proposed for image completion, and *shift-maps* [PKVP09], for image retargeting and reshuffling. However, we show that a straightforward extension of these image methods to videos does not lead to coherent results (Sec. 5). We transfer the correspondence-map concept to video by encoding the trade-off between spatial and temporal coherence in the energy function, and by handling the special conditions at the occlusion boundary. The resulting energy is suitable for global optimization using the expansion move algorithm and graph cuts [BK04, KZ04, BVZ01], and we show that it generalizes well across sequences with constant parameters.

Additionally, we propose an interface to make effective use of user input (Sec. 4). First, computation times can be drastically reduced by incorporating simple user strokes marking the track of occluded objects to constrain the search space. Second, the user can refine the result by approximately specifying the desired source and target regions.

State-of-the-art completion approaches exploit redundancies in videos [WSI07] and images [PKVP09, BSFG09]. We performed an extensive evaluation of the direct application and the application of simple adaptations of these approaches to video completion. Shortcomings of these techniques helped shape our approach, and we discuss this in Sec. 3.3. The state-of-the-art method for non-parametric video completion of Wexler et al. [WSI07] pioneered exploiting space-time redundancies by defining a global energy that is locally optimized. When compared with this method, our approach leads to improved results that are more spatially and temporally coherent (Sec. 5). We achieve this through a newly designed energy function and an effective MRF optimizer that maximizes coherence directly in the offset-space (as opposed to in the image space).

In summary, our contributions are:

- A new energy functional for the problem of video completion that is shown to improve visual coherence when compared to state-of-the-art non-parametric methods, and that performs well on diverse scenes using a constant set of parameters; and
- a system and user interface for interactively guiding the completion process such that computation times can be drastically reduced and errors quickly remedied.

Compared with previous methods, we demonstrate our framework on more challenging scenes (Sec. 5), which feature multiple occlusions, complex motions, and high-resolution objects, where mistakes are more noticeable than in their low-resolution counterparts.

2. Related Work

Existing non-parametric video completion algorithms can be broadly classified into two categories: *global* and *local* methods. The first category employs a *global* energy minimization framework where the requirements of video completion plus any a priori knowledge are encoded into a single energy functional whose minima provide globally consistent solutions. As these approaches aim to distribute the error across the domain, they can be applied even when the hole is large in space and in time. Nevertheless, due to its high computational complexity, the optimization is commonly approached through the use of approximate methods. In the category of *local* methods, information is greedily propagated from outside the hole into the missing region. This strategy generally results in faster algorithms. However, as local information propagation does not guarantee global consistency, these methods are less suitable for large spatio-temporal holes.

Wexler et al.'s method [WSI07] lies in the first category. A spatio-temporal patch is sampled from every observed pixel in the input video. The collection of these patches constitute a database reflecting the statistics of the video. Completion is performed by greedily assigning to each missing pixel the most likely color among those patches from the database that are closest to the current solution. The joint configuration of these assignments minimizes a predefined energy functional based on the (dis-)agreement of overlapping patches, therefore casting video completion as a global energy minimization. The method of Shen et al. [SLCF06] tries to retain the advantages of global approaches while reducing computational complexity. They track every pixel in the occludee through the video such that during the energy minimization stage the search space for each pixel is reduced from 3D to a

2D manifold. Under this simplification, the method can handle only pure translation of rigid objects or periodic motions.

Our global method builds upon the *shift-map image editing* [PKVP09] framework, which produces a vector field or *shift-map* that assigns an offset to every pixel such that the resulting field minimizes a task-dependent energy functional. For performing image completion, they constrain the offsets assigned to missing pixels such that they point to pixels outside the hole from which colors should be copied. They optimize a MRF that minimizes the discrepancy between the (original) neighborhoods of adjacent pixels. Hu and Rajan's *hybrid shift map* [HR10] applies the concept of shift-map to the video domain. For video retargeting, they obtain a coherent mapping from the original video to a domain of different resolution by maximizing temporal consistency in addition to spatial consistency. Offsets along the time axis are not allowed nor required during the retargeting. However, completing video parts by simply extending shift-maps to allow temporal offsets can produce spurious artifacts since, in general, spatial and temporal dimensions have different characteristics (Fig. 4). Any energy function has to be re-designed to reflect these differences (Sec. 3).

Patwardhan et al. [PSB05] propose a local approach. They assign a priority to every hole pixel based on the local amount of undamaged pixels and on the presence of structure at the current hole boundary. Proceeding by highest priority, the method copies those patches that best match the context of the pixel of interest. This algorithm was later improved to handle camera motions parallel to the image plane [PSB07]. While this algorithm enables fast completion, in general the results are not guaranteed to be globally coherent. In [STH09], this work is extended to handle general camera motions and remove temporal discontinuities.

A more indirect approach is motion transfer. The idea is to derive a motion field for the hole by gradually propagating motion vectors [MOG*06] or by using motion patch similarities [SMTK06]. The motion field is used to propagate pixel values from outside the hole into the missing region. These approaches allow completion only over a relatively small number of frames. Completion of large time intervals is not straightforward as pixel propagation tends to suffer from smoothing artifacts.

Several other video completion techniques rely on object-based hole-filling. The method by Venkatesh et al. [VCZ09] tracks and segments the occluded object. They construct a database of segmented frames where the occluded object is fully visible. Using dynamic programming, the holes are completed by aligning frames in the database to the partially or fully occluded frames in the hole. This requires segmentation to be very accurate and motions to be mostly cyclical for each occluded object. This idea is extended by Ling et al. [LLS*09]. The contours of the object of interest are estimated by using motion information. The contours are then used to retrieve the object frames from a database using an approach similar to Venkatesh et al. [VCZ09]. To find database frames despite differences in posture, query postures are synthesized based on local segments of the object. Similarly, the technique by Jia et al. [JTWT06] assumes a periodic motion for the occluded objects. After segmenting them, the completion problem is cast as a warping and aligning of the object's visible trajectory with that of its occluded views. Object-based systems demonstrate plausible completions for certain categories of moving objects, especially humans. However, they either impose an explicit class of possible motions (e.g., cyclic) [VCZ09, JTWT06] or require the motion to be simple such that a dense sampling of postures is feasible [LLS*09]. Furthermore, by design, the completion of objects is performed independently of the background. Consequently, the blending between the completed foreground and the background can look unnatural. In contrast, non-parametric methods [PSB07, JHM05, SMTK06, WSI07, STH09] do not suffer from these limitations but cannot take advantage of model-based priors.

Lastly, Shih et al. [STTHY08] approach the related problem of modifying the motion of people in videos. Their inpainting component is a 3D extension of the exemplar-based image inpainting algorithm by Criminisi et al. [CPT04]: The boundary of the hole is filled by searching for the locations that best match the texture and motion contexts. To reduce the time complexity of the search process, each person is tracked in the video, and a *skeleton* is computed for each. By propagating the skeleton into the corresponding object hole and establishing the correspondences between these figures throughout the frames, the search space is reduced. However, they expect the motion to be cyclic and assume that a 2D skeleton model can be reconstructed from the input video.

3. Automatic Object Completion

The input to our algorithm is a video sequence or *video volume* $I(x,y,t)$ that shows several dynamic scene elements, e.g., people which occlude each other for certain periods of time. In the first step, we construct a mask for the scene element O_r to be removed. To this end, we use *Video SnapCut* [BWSS09]. O_r leaves a spatio-temporal hole $\Omega(O_r)$ in the video volume (Fig. 2) that needs to be filled, i.e., the background (static and dynamic), and every moving object $\{O_{c,i},\ldots,O_{c,m}\}$ which are occluded by O_r have to be completed in a spatio-temporally coherent manner. We achieve this by first inpainting the background behind O_r, and then completing the m dynamic occlusions remaining in the video after the background is inpainted.

For each missing pixel in $\Omega(O_r)$, we attempt to find a spatio-temporal displacement (or offset) that points to another unoccluded pixel from which to copy the missing color [PKVP09]. For a given video I and a hole $\Omega(O_r)$, we construct a *shift-volume* $M(x,y,t) = (d_x,d_y,d_t)$ for $(x,y,t) \in \Omega(O_r)$. The output video R is constructed by assigning to each location $(x,y,t) \in \Omega(O_r)$ the color values of its

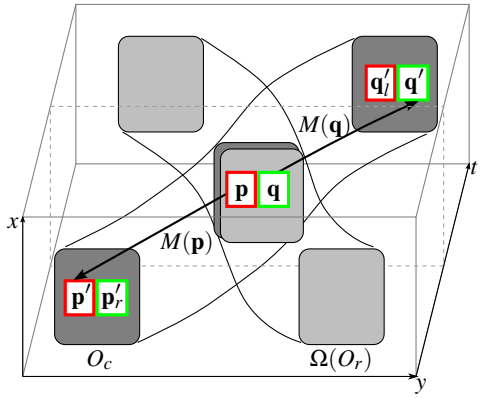


Figure 2: *Video volume inpainting. A pair of missing pixels* $\mathbf{p},\mathbf{q}$ *inside the spatio-temporal hole* $\Omega(O_r)$ *can be replaced by pixels* $\mathbf{p}' \equiv \mathbf{p}+M(\mathbf{p})$, $\mathbf{q}' \equiv \mathbf{q}+M(\mathbf{q})$ *outside the hole provided that the appearance of their neighborhoods is consistent, i.e., that the color and gradient mismatch between the pairs* $(\mathbf{p}',\mathbf{q}'_l)$ *and* $(\mathbf{p}'_r,\mathbf{q}')$ *is sufficiently low. For an inpainting to be satisfactory, such neighborhood consistency has to be maintained for all adjacent pixels in* $\Omega(O_r)$.

shifted location, i.e., $R(x,y,t) = I((x,y,t)+M(x,y,t))$. We find M by minimizing a global energy functional that models the trade-off between two competing objectives: (a) The shifts should be as homogeneous as possible, which implies that the corresponding region is obtained from a spatio-temporally coherent segment in the video, and therefore appears natural; and (b) the boundaries between different homogeneous regions in the shift-volume and across the boundary of the hole should not look objectionable.

3.1. Energy Functional

We use vectorial notations for denoting indices of spatio-temporal pixels, i.e., pixels $\mathbf{p},\mathbf{q} \in \mathcal{I} \subset \mathbb{Z}^3$, where $\mathcal{I}$ is the index set for the entire input video. Let $\Phi \equiv \mathcal{I} \setminus \Omega(O_r)$ be the unoccluded region. The shift-volume M is obtained as a minimizer of the energy functional

$$\mathcal{E}(M) = \sum_{\mathbf{p}\in\Omega(O_r)} \sum_{\mathbf{q}\in\mathcal{N}(\mathbf{p})} \mathcal{V}_{\mathbf{p},\mathbf{q}}(M(\mathbf{p}),M(\mathbf{q})), \qquad (1)$$

with the condition that all offsets should point to outside of the hole, i.e., $\forall \mathbf{p} \in \Omega(O_r) : \mathbf{p}+M(\mathbf{p}) \in \Phi$. Here, $\mathcal{N}(\mathbf{p})$ denotes the set of pixels adjacent to $\mathbf{p}$ in a 26-neighborhood system. The pair-wise smoothness term $\mathcal{V}$ represents the cost of discontinuities in the shift volume. We measure the discrepancy of the corresponding pixel values using the distances of the color and gradient values within the local neighborhoods [KSE*03] of $\mathbf{p}$ and $\mathbf{q}$:

$$\mathcal{V}_{\mathbf{p},\mathbf{q}}(M(\mathbf{p}),M(\mathbf{q})) = \begin{cases} 0 & \text{if } M(\mathbf{p}) = M(\mathbf{q}), \\ h(\mathbf{p},\mathbf{q})\gamma(\mathbf{p},\mathbf{q}) & \text{otherwise,} \end{cases} \qquad (2)$$

where

$$\begin{aligned} h(\mathbf{p},\mathbf{q}) = \Big(\| & I(\mathbf{p}+M(\mathbf{p})) - I(\mathbf{p}+M(\mathbf{q}))\|_2^2 + \\ \| & I(\mathbf{q}+M(\mathbf{p})) - I(\mathbf{q}+M(\mathbf{q}))\|_2^2 \Big)^{\psi} + \\ \beta \Big(& \|\nabla I(\mathbf{p}+M(\mathbf{p})) - \nabla I(\mathbf{p}+M(\mathbf{q}))\|_2^2 + \\ & \|\nabla I(\mathbf{q}+M(\mathbf{p})) - \nabla I(\mathbf{q}+M(\mathbf{q}))\|_2^2 \Big)^{\psi} + \lambda, \end{aligned} \qquad (3)$$

and β is the weight balancing the contribution of gradient and color values, which is fixed to $(2\sqrt{2})^{-1}$ throughout all experiments. This value is based on the observation that the range of gradient differences is twice as large as that of the pixel differences (hence the factor $\frac{1}{2}$), and their variance, assuming linear camera response, is also twice as large (hence the factor $\frac{1}{\sqrt{2}}$). To account for noise in the differences, we add a small constant $\lambda = 0.1$ to every term.

The exponent ψ in Eq. (3) is fixed to $\frac{1}{2}$. This introduces robustness against outliers, in contrast to the L_2 square regularizer used by Pritch et al. [PKVP09]. In addition, this choice of parameters ψ, λ ensures that h is a metric, and leads to a *sub-modular* or *regular* functional [BVZ01]. This generates solutions near to the global minima with an optimization using graph cuts: Theorem 6.1 of [BVZ01] states that the energy of the final solution is bounded by a constant multiple of the minimum, and this is a desirable behavior of the optimizer. This is supported by our experiments (Sec. 5).

The role of the weight function γ is twofold. First, it balances the cost of spatial and temporal discrepancies, which, in general, have different characteristics. The second role is to enforce the consistency of color values of the pixels which are closer to the boundary [WSI07]. A uniform enforcement of pairwise similarities $\mathcal{V}$, i.e., without weighting by γ, can result in undesirable artifacts in results. For a large hole, the ratio between the hole boundary and the hole volume is large. In this case, without γ, the penalty for inconsistencies occurring inside the hole can override the penalties at the boundary, and accordingly, it can result in a completely uniform shift-map. In general, we want to emphasize the consistency at the boundary more than in the hole interior so the information of the visible parts around the hole is *propagated* into the hole.

We construct γ such that when the hole *depth* of a pair $(\mathbf{p},\mathbf{q})$ is smaller than the depth of pair $(\mathbf{r},\mathbf{s})$ the corresponding weight $\gamma(\mathbf{p},\mathbf{q})$ is sufficiently bigger than $\gamma(\mathbf{r},\mathbf{s})$. The depth $d_0(\mathbf{p},\mathbf{q})$ is defined as $d_0(\mathbf{p},\mathbf{q}) = \frac{1}{2}[d(\mathbf{p},\partial\Omega(O_r)) + d(\mathbf{q},\partial\Omega(O_r))]$, where the distance $d(A,\mathcal{B})$ between a point A and a set $\mathcal{B}$ is defined as the minimum of the *Chebyshev* distance (corresponding to a 26-neighborhood system) between A and the elements of $\mathcal{B}$. Based on d_0, we partition the set of pairs $\{(\mathbf{p},\mathbf{q})\}$ in the hole so that each partition P_i consists of pairs which have the same distance. In addition, we arrange the partitions in the order of increasing distances (i.e., $d_0(\mathbf{p},\mathbf{q}) < d_0(\mathbf{r},\mathbf{s})$ for $(\mathbf{p},\mathbf{q}) \in P_i$, $(\mathbf{r},\mathbf{s}) \in P_j$, and $i < j$). An intermediate weight function γ_1 is defined

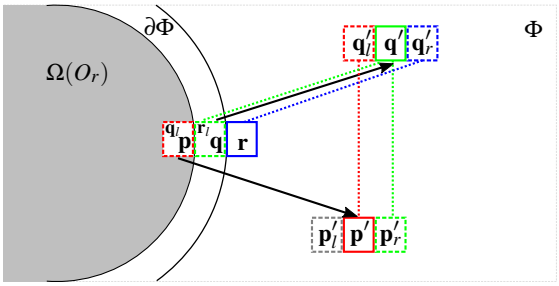


Figure 3: *In addition to finding offsets for pixels* $\mathbf{p} \in \Omega(O_r)$*, we allow shifts for pixels* $\mathbf{q} \in \partial\Phi$ *adjacent to the hole. The dotted lines denote the pixel differences involving* $\mathbf{q}$ *that are evaluated in Eq. (1), namely* $\mathcal{V}_{\mathbf{p},\mathbf{q}}(M(\mathbf{p}),M(\mathbf{q}))$ *and* $\mathcal{V}_{\mathbf{q},\mathbf{r}}(M(\mathbf{q}),\mathbf{0})$*. By allowing boundary pixels* $\mathbf{q}$ *to shift, we ensure that their neighbors (which could be inside the hole, e.g.* $\mathbf{q}_l$*) are always defined.*

based on the recurrence relation $\gamma_1(P_{i+1}) = \frac{1}{2}\gamma_1(P_i)\frac{|P_{i+1}|}{|P_i|}$, with $\gamma_1(0) = 1$, where $|P_i|$ is the size of P_i. This guarantees that the distinct partitions are sufficiently separated, i.e. $2\gamma_1(P_i)|P_i| \geq \gamma_1(P_j)|P_j|$ if $i < j$. The final weight function γ is obtained as

$$\gamma(\mathbf{p},\mathbf{q}) = \begin{cases} \alpha\gamma_1(\mathbf{p},\mathbf{q}) & \text{if } \mathbf{p}-\mathbf{q} = (0,0,\pm 1) \\ \gamma_1(\mathbf{p},\mathbf{q}) & \text{otherwise,} \end{cases} \quad (4)$$

where α is a constant which balances the contributions of incoherences occurring in temporal and spatial dimensions. We fix α to be $\frac{8}{18}$, which is the ratio between the number of neighbors in the same frame and the number of neighbors in adjacent frames. This reflects the fact that temporal incoherence should be at least as important as spatial incoherence.

Pixels close to the boundary of the hole deserve special treatment (Fig. 3). As currently specified, $\mathcal{V}$ is undefined for pixels $\mathbf{q} \in \partial\Phi$. Since $\mathbf{q}$ is outside the hole, it is not allowed to shift by the constraint $M(\mathbf{q}) = \mathbf{0}$. In this case, the values $I(\mathbf{p}+M(\mathbf{q}))$ and $\nabla I(\mathbf{p}+M(\mathbf{q}))$ for neighbors $\mathbf{p}$ of $\mathbf{q}$, $\mathbf{p} \in \Omega(O_r)$ would be drawn from inside the hole, where they are undefined. We solve this by allowing pixels at the boundary to be shifted such that their neighbors become well defined.

While our energy functional takes inspiration from several existing methods, including [PKVP09] and [WSI07], it differs in important ways. In contrast to [PKVP09], an L_2 penalizer allows completions under larger differences between an object's exemplar and its missing appearance. Additionally, the inclusion of the weighting function γ ensures the correct direction of information flow. As we show in Sec. 5, a naïve application of shift-map to video without the weighting function leads to inferior results. The use of γ is inspired by [WSI07], but differs in that we make distinction between the temporal and spatial coherences (see Eq. (4)). Additionally, we explicitly enforce the shifting of connected segments, which is a natural choice since connected segments outside the hole are *observed* and therefore they are themselves coherent. The energy functional (Eq. 2) prefers shifting as large segments as possible while keeping a consistent appearance across different segments. In Fig. 4 and Sec. 5.1, we demonstrate how these differences in the energy functional and the optimization strategy can lead to significant differences in the final results.

3.2. Optimization

The energy functional in Eq. (1) is non-convex and finding a global minimum is difficult. Instead, we find an approximate solution using graph cuts [BK04, KZ04, BVZ01] where each individual node corresponds to a pixel in $\Omega(O_r) \cup \partial\Phi$, and the set of potential labels corresponds to the set of possible offsets within the video volume.

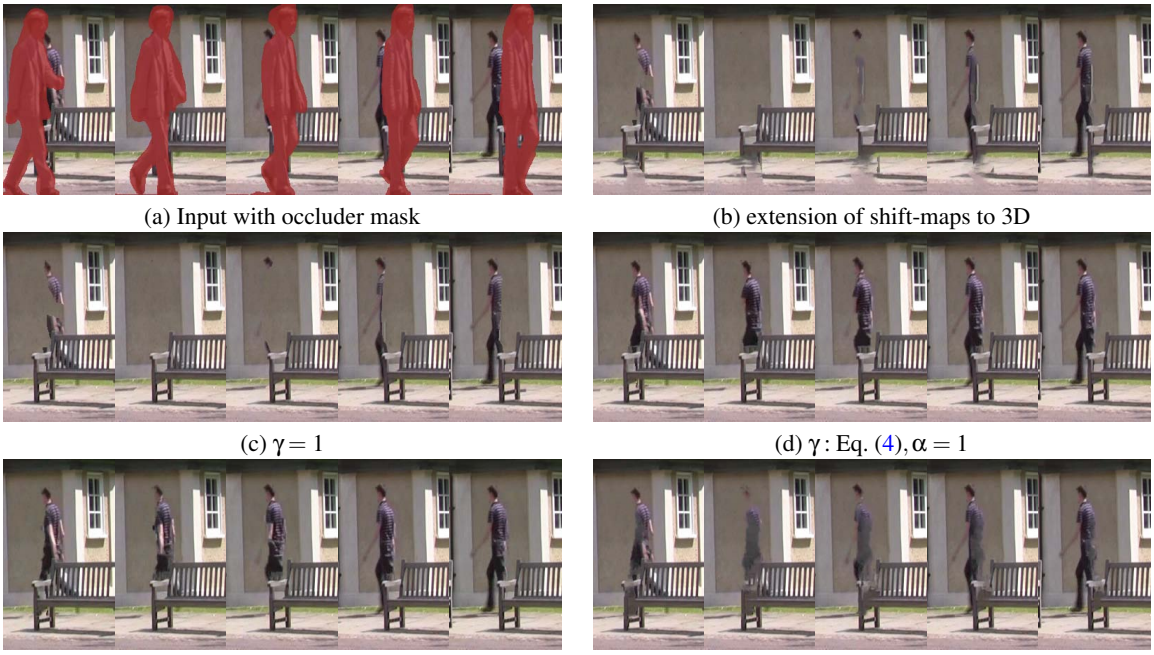
Directly minimizing the energy functional using graph-cuts is still a challenging problem due to the very large size of the label space. Similar to [PKVP09], we adopt a multi-resolution approach. First, a video pyramid is generated by reducing the spatial resolution by half until the optimization of Eq. (1) becomes tractable. Masks are down-sampled in a conservative way such that holes in finer levels remain holes in coarser levels. Down-sampling is not performed along the time axis as it can introduce temporal discontinuities.

Once the solution (the shift-volume) is found at the coarsest scale, it is up-sampled to an initial guess for the next higher-resolution level using nearest neighbors interpolation. The magnitudes within the shift-volume are doubled to match the higher resolution. This step is repeated until we reach the original resolution. On the lowest pyramid level, the size of the label space is $(\frac{2w}{k}-1)(\frac{2h}{k}-1)(2t-1) \approx \frac{4wht}{k^2}$, where w,h,t are, respectively, the width, height, and length of the video and k is the reduction factor of the coarsest pyramid level. We set k such that the number of nodes in the graph is smaller than 100^3. In this way, the optimization remains feasible on standard computing hardware. For the initial background completion, the size of the label space is $(2t-1)$ as spatial shits are not allowed. On all levels above the coarsest, only small shifts relative to the initial estimate are examined. In our implementation, we use three relative shifts $(-1,0,1)$ in each spatial and temporal coordinate.

3.3. Design Validation

Figure 4-e shows that the distance weighting function γ, as well as the proper setting of the relative importance of time α are important for obtaining high-quality completion results. If the time domain is given a higher importance ($\alpha = \frac{8}{18}$) but the distance-to-boundary weighting is not used ($\gamma_1 = 1$), completions that (erroneously) restore only the background receive a low penalty (Fig. 4-c). In this case, the interior of the hole is coherent, but the discontinuities at the hole boundary are not weighted sufficiently highly.

Conversely, if we enable the distance weighting γ_1 but do



(a) Input with occluder mask (b) extension of shift-maps to 3D

(c) $\gamma = 1$ (d) γ : Eq. (4), $\alpha = 1$

(e) γ : Eq. (4), $\alpha = \frac{8}{18}$ (proposed) (f) γ : Eq. (4), $\alpha = \frac{8}{18}$, $\psi = 1$

Figure 4: *Effect of distance weighting γ and temporal weight α: (a) Input video and overlaid mask of O_r. (b) Straightforward extension of 2D shift-maps to 3D. (c) No distance weighting ($\gamma = 1$). (d) When using γ as defined in Eq. (4) and $\alpha = 1$, the result is spatially consistent but shows slight temporal misalignments. (e) Proposed method: With γ as defined in Eq. (4) and $\alpha = \frac{8}{18}$, the balance between spatial and temporal consistency is kept. (f) Same as (e) but using a quadratic penalizer as in (b).*

not increase the importance of temporal mismatches ($\alpha = 1$), we receive spatially coherent (for individual frames) but temporally incoherent inpainting results. This effect is seen in the incorrectly located head in Fig. 4-d (third frame). The importance of properly considering the time domain is also documented in the literature on visual perception. Wandell [Wan95] states that the human sensitivity to temporal contrast changes peaks roughly between 5 Hz to 10 Hz. That is, temporal aliasing at about a third of the frame rate is very objectionable to the viewer. If the time domain is not re-weighted, we can suffer temporal aliasing in that range.

Lastly, the irregular variant of our proposed energy, which we obtain by using a quadratic penalizer ($\psi = 1$), does not produce the desired outcome, and it leads to over-smoothed, washed out completions (Fig. 4-f). In contrast, our proposed energy function encodes the elements – distance weighting, trade-off between spatial versus temporal mismatches, robust penalizer – necessary to achieve high quality results.

4. User Guided Completion

Naturally, inpainting results can be improved by exploiting information about the semantics of the scene. This improvement can be in terms of speed and of increasing the chances of a successful inpainting. We developed an interface to provide semantic constraints that help reduce the run-time and allows the user to correct remaining inpainting errors.

To make the optimization feasible on high-resolution videos, we need to reduce the search space (i.e., the set of candidate sources) of an occluded object $O_{c,i}$. Assuming a stationary camera, we perform foreground thresholding using an estimated background model [GSL08] in order to prune out irrelevant background regions from the search space. To further restrict the search space, use the observation that valid sources for completing an object lie within a small video volume centered around the object's trajectory [JHM05]. Given that the user can easily discern the location of the occluder and occludee in crowded scenes, we provide an interface to quickly specify the trajectories of occluded objects (see supplementary video). In the interface, the user masks the occluded objects $O_{c,i}$ on xt- and ty-projections of the video volume (Fig. 5). From these masks, we determine a bounding box for each occluded object on every frame, which we use to constrain the space of potential inpainting sources for the object. We apply this user-guided tracking in all results presented in Sec. 5, except for the *beach-umbrella* and *duo* sequences.

While our algorithm produces better inpainting results than existing approaches (Sec. 5.1), it can fail to produce

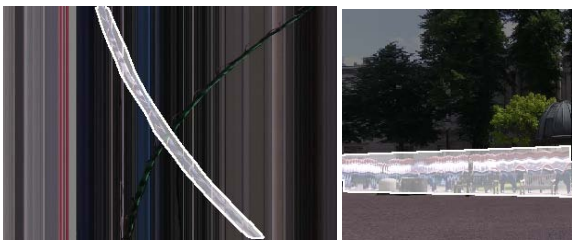
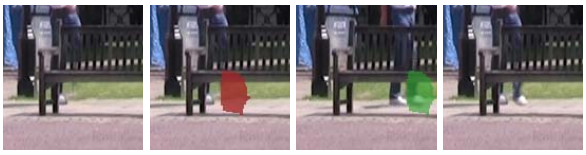


Figure 5: *Interface for occludee tracking: To reduce the computation time, we restrict the space of possible shifts to the region spanned by the occluded object O_c. In highlight is the coarse mask drawn by the user on top of a xt projection (left) and a ty projection (right) of the input video.*



(a) (b) (c) (d)

Figure 6: *User-assisted refinement: (a) After automatic inpainting the leg of the person was incorrectly completed. (b, c) The user marks the target region to be refined (red), and marks a suitable source region in the video volume (green). (d) After computing a solution on the constrained space, the error is corrected.*

plausible inpaintings due to its non-parametric nature. Fortunately, such errors usually materialize in a spatio-temporally confined region of the video. The missing leg on the inpainting in Fig. 6 illustrates this situation. To assist the inpainting method in correcting such artifacts, we provide a tool for the user to add semantic understanding of the scene to the algorithm (see supplementary video). If the user constrains the search space to only relevant spatio-temporal source regions, such as regions where a leg is seen in the above example, the inpainting algorithm is less likely to select incorrect (but lower energy) sources to complete the missing region. Such user guided refinement has been previously demonstrated for images [PPP11, BSFG09]. All results shown in this paper and in the supplemental video (except Fig. 6) were produced without user-assisted refinement.

5. Results

We tested our algorithm on six videos, which we will referred to as *beach-umbrella, park-groundtruth, park-simple, park-complex, duo* and *museum* (see Fig. 1, 7-10, and supplementary video). The *beach-umbrella* sequence (Fig. 10-top), introduced in [WSI07], is of comparably low resolution, with occludee sizes of around 25×50 pixels. The remaining sequences come from new footage shot in Full HD, with occludee sizes going up to $\sim 384 \times 512$. To our knowledge, no previous approach has demonstrated completion on videos of the complexity of *park-complex* (Fig. 7) or *museum* (Fig. 9). For such high-resolution scene elements, any completion artifact will be more noticeable. In Sec. 5.1, we compare our method with previous work using the simpler *umbrella*, *park-groundtruth*, and *park-simple* sequences.

In the *duo* sequence, we remove two pedestrians that occlude two music performers standing in front of a reflective surface (Fig. 8). During the occlusion, the performers display repetitive hand movements that need to be recovered. In the background, the pedestrians also occlude the reflections of other moving objects in the scene. Our method successfully completes the dynamic foreground and background scene elements, and produces a very plausible result.

In *park-complex* (Fig. 7), we remove a person who occludes seven other people, each displaying different behaviors such as sitting, standing, and walking towards, away from, and parallel to the camera. Cases 1, 6, 7 contain people that are non-periodically moving: In case 7 the person stands up while being half-occluded, and in cases 1, 6 they start walking right after the occlusion. In case 3, the person is turning on the spot and walking away from the camera during the occlusion. Case 2 is particularly challenging: While the body and motion of the person is realistically completed, our method is challenged by the fact that he raises his arm during the occlusion and that this type of motion is not seen elsewhere in the video. Case 5 is another demanding example where the person is being occluded by an static obstacle (a bench) at the same time it is occluded by the object we mark for removal. Whereas many occlusions are successfully completed, we observed that most completion artifacts occur when the behavior or appearance of the occluded object is different during occlusion and disocclusion, thus breaking the method's assumptions.

The *museum* sequence (Fig. 1, 9) sequence is the most challenging set: We remove a person who occludes eight other people, several in high-resolution (up to 384×512), walking over a specular floor. The people are located at different distances from the camera, and they show different types of motion such as standing and moving parallel to and away from the camera plane. Our method produces satisfactory completions for the two large occlusions in cases 3 and 4, despite their high-resolution. Furthermore, the algorithm successfully completed the reflections on the floor, which is specially noticeable in cases 2 and 4. In case 2, the person is walking away from the camera but the method nevertheless accomplishes a coherent completion with only slight discontinuities visible when slowing the video down. Given that the person in this case is walking away from the camera, there is perspective foreshortening. Accordingly, there are no exactly matching examples in the video that can be used to fill in the occluded region.

In all these challenging cases, the resulting video comple-

tion looks convincing (see supplementary material). A close look at the individual frames can reveal artifacts which are less noticeable when playing the video. Most of these artifacts can be removed with our approach with user-guided refinement. Fig. 6 shows an example where the leg of a person was not correctly filled in. The user specified a region containing an unoccluded view of the leg, and the algorithm recomputed the result using this constraint search space. With this simple user interaction and after a few minutes of computation, the error was corrected.

5.1. Comparison to Related Approaches

We implemented the method of Wexler et al. [WSI07] as an alternative state-of-the-art non-parametric method, and an extension of the 2D approach of Pritch et al. [PKVP09] to 3D (i.e., augmenting the 2D image domain to 3D by simply adding the temporal dimension). We compared the resulting completions on the sequences *beach-umbrella*, *park-groundtruth*, and *park-simple* (Fig. 10). In the first sequence, a still umbrella occluding three walking people is removed, in *park-groundtruth*, we simulated a moving human occluder, and in *park-simple*, we remove a pedestrian that only occludes one other walking person.

On the low resolution *beach-umbrella* sequence, all the three methods produced coherent results (Fig. 10-top).

On *park-groundtruth*, the result of a naïve extension of Pritch et al.'s method to 3D introduced temporally incoherent structures at the occlusion boundary (Fig. 10-c-middle). Without the weighting function, the cost of introducing such discontinuities at the boundary are offset by the low cost of inpainting at a coherent background inside the hole. This supports the introduction of γ in our energy functional (Sec. 3). The result of Wexler et al.'s method is satisfactory but introduces artifacts on some thin structures: A leg and a part of the arm are inpainted with background (Fig. 10-b-middle). Our algorithm convincingly reconstructed the motion of the occluded person (Fig. 10-d-middle).

In the *park-simple* sequence (Fig. 10-bottom), the result of Pritch et al. shows that the person was completed using a smooth transition to the background behind. This underlines the importance of properly balancing spatial and temporal coherence to prevent these motion discontinuities. The result from Wexler et al. shows a missing thin structure (an arm). In comparison, our result is more coherent, although in some frames the hand was not properly completed. However, this artifact is much less noticeable than those present in the other two results. Please refer to the supplemental video to fully appreciate these remarks.

5.2. Parameter Selection and Timings

For all our results, we use the same set of parameters: $\alpha = \frac{8}{18}$, $\beta = (2\sqrt{2})^{-1}$, and $\psi = \frac{1}{2}$ (see Sec. 3.1 and Sec. 3.3 for a discussion on these specific choices). This shows the robustness and stability of our approach across different types of scenes. We limited the number of iterations of the expansion-move algorithm to five. Observing that occludees are not rapidly approaching or moving far away from the camera, we restricted the possible values of M along the y-axis to the interval $[-16, 16]$. The latter was introduced to speed up the experiments and is not a requirement of the method.

When computing solutions for individual occlusions in parallel on a Xeon X5560 CPU, it took between 11 hours (*beach-umbrella*) and 90 hours (*museum*) to process an entire video with all containing occlusions. We abstain from performing a direct run-time comparison with Wexler et al. as the two implementations run on different platforms (C++ vs. Matlab). It takes around 30 minutes to interactively create the mask of an occluder using *Video SnapCut* [BWSS09], and less than a minute to track the path of each occludee. Such tracking usually reduces the reference region to less than a quarter of the original size. Recall that the size of the label space is $N \approx 4wht/k^2$ (Sec. 3.2). As the run-time complexity is $O(n^3N)$, where n is the number of missing pixels ($n << N$), a reduction by a fourth on the width and height implies a 16x speed up of the running time.

Figure 7: *Completion of the* park-complex *video: (top) Sample frame from six of the occlusions occurring in the input video. (bottom) Result of our completion algorithm.*

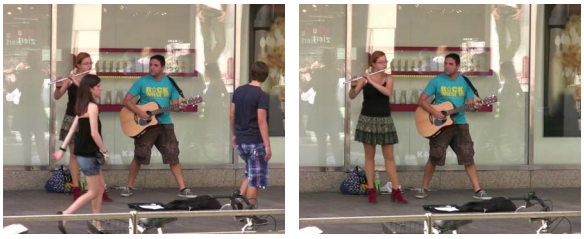


Figure 8: *Completion of the* duo *sequence with two occluders and two occludees: (left) sample frame, (right) our result.*

Figure 9: *Completion of the* museum *video: (top) Sample frames from the four largest occluded moving objects. (bottom) Our result.*

Figure 10: *From top to bottom: the* beach-umbrella*, the* park-groundtruth*, and the* park-simple *sequences. (a) Input frame, (b) result by Wexler et al., (c) naïve extension of Pritch et al. to 3D, (d) our result. In the umbrella sequence all three approaches produce plausible results. Our approach outperforms the others on the more challenging park- sequences. Our improvement is better appreciated when watching the supplemental video, where the artifacts of the other approaches are more obvious.*

6. Limitations and Future work

By design, the synthesized color values filling the holes have to come from unoccluded regions found in the the input. Accordingly, the system might fail if at the time of occlusion the object had a different appearance due to lighting changes, motion, or deformation. To our knowledge, no approach that is designed to work on general scenes and refrains from strict model assumptions about scene content can fully-automatically handle such cases. One possibility to overcome this limitation is to use additional user input: e.g., the user could explicitly mark the spatio-temporal regions to use as a example, and we support this case. For specific applications, e.g., inpainting humans, several methods have resorted to using stronger model assumptions, such as motion and shape models that could be tracked and used to constrain the sources in a semantically meaningful way.

Currently, we do not use speed or acceleration information, so we cannot explicitly encourage preservation of these properties. The energy function could be extended to enforce similarity between source and target using higher order derivatives. However, computing higher order derivatives is potentially unstable and further increases computation time.

Although we experimented with videos from static cameras, the proposed method could be extended to support moving cameras. However, our main assumption (i.e., missing data being available elsewhere in the video) is harder to satisfy in this setting as both the motion of the object and of the camera has to be matched. This requirement is especially difficult to satisfy with hand-held cameras, where, in addition, motion blur artifacts occur. These effects could be lessened by performing video stabilization and deblurring.

Our energy function as defined in Eq. (1) is not scale invariant. This makes it difficult to cope with objects moving away from or towards the camera in cases where the missing scale was not observed. We can handle such situations provided that the scale does not change significantly during the occlusion, as demonstrated in Fig. 9. This limitation could be overcome by allowing shifts along the video scale space. Furthermore, to extend our approach to sequences with noticeable illumination variations, one could perform completion in the gradient domain only and use a 3D Poisson equation solver [Rob99] to recover the final video.

Even though the run time of our algorithm can be long, especially on high-resolution sequences, the quality of the results is high and many remaining artifacts can be efficiently remedied with simple user guidance. We believe that this is a step ahead of the standard practice of performing manual inpainting. Professional video or animation artists are thus free to do more creative work while our algorithm runs in the background. We plan to investigate fast local solvers, such as PatchMatch [BSFG09], to provide results at close to interactive rates, albeit potentially sacrificing quality.

In the future, we would like to use video inpainting as a building block for other applications. For instance, one could

inpaint every occlusion of every person in a video sequence, such as the *museum* clip, to create loop-able tracks for every person. This would enable the creation of novel infinitely long sequences by overlaying the tracks in varying order.

7. Conclusion

This paper presents a non-parametric method for video completion based on the estimation of a shift-volume. We assign every pixel in the missing region the color of an unoccluded observation, such that the result looks natural and consistent with the hole boundary. By encoding these requirements in a global energy functional, video completion can be regarded as selecting for every pixel the most compatible source outside the hole, such that the resulting energy is minimized. By following the trajectory of each occluded object with the help of user-interaction, we reduce the search space from the entire video volume to a tracking window around the occluded objects, leading us to a computationally feasible formulation of the problem. We show that interactive search space reduction can also be applied to refine the automatic completion results. We have successfully applied our system to more challenging sequences than those previously shown in the literature. The results indicate that the performance of the proposed method is significantly better than previous state-of-the-art methods.

Acknowledgements Significant thanks to Yonatan Wexler for his generous and patient help, and to Yael Pritch for answering our questions. This work was supported by BBC R&D, and by the EngD VEIV Centre at UCL.

References

[BK04] BOYKOV Y., KOLMOGOROV V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE TPAMI 26*, 9 (2004), 1124–1137. 2, 5

[BSFG09] BARNES C., SHECHTMAN E., FINKELSTEIN A., GOLDMAN D. B.: PatchMatch: a randomized correspondence algorithm for structural image editing. *ACM Trans. Graphics (Proc. SIGGRAPH) 28*, 3 (2009), 24:1–24:11. 2, 7, 9

[BVZ01] BOYKOV Y., VEKSLER O., ZABIH R.: Fast approximate energy minimization via graph cuts. *IEEE TPAMI 23*, 11 (2001), 1222–1239. 2, 4, 5

[BWSS09] BAI X., WANG J., SIMONS D., SAPIRO G.: Video snapcut: robust video object cutout using localized classifiers. *ACM Trans. Graphics (Proc. SIGGRAPH) 28*, 3 (2009). 3, 8

[CPT04] CRIMINISI A., PÉREZ P., TOYAMA K.: Region filling and object removal by exemplar-based image inpainting. *IEEE TIP 13*, 9 (2004), 1200–1212. 3

[DSC03] DEMANET L., SONG B., CHAN T.: Image inpainting by correspondence maps: a deterministic approach. *Applied and Computational Mathematics 1100* (2003), 217–50. 2

[GBI09] GLASNER D., BAGON S., IRANI M.: Super-resolution from a single image. In *Proc. ICCV* (2009), IEEE, pp. 349–356. 2

[GSL08] GRANADOS M., SEIDEL H.-P., LENSCH H. P. A.: Background estimation from non-time sequence images. In *Proc. Graphics Interface* (2008), pp. 33–40. 6

[HR10] HU Y., RAJAN D.: Hybrid shift map for video retargeting. In *Proc. IEEE CVPR* (2010), IEEE, pp. 577–584. 3

[JHM05] JIA Y.-T., HU S.-M., MARTIN R. R.: Video completion using tracking and fragment merging. *The Visual Computer 21*, 8-10 (2005), 601–610. 3, 6

[JTWT06] JIA J., TAI Y.-W., WU T.-P., TANG C.-K.: Video repairing under variable illumination using cyclic motions. *IEEE TPAMI 28*, 5 (2006), 832–839. 3

[KSE*03] KWATRA V., SCHÖDL A., ESSA I. A., TURK G., BOBICK A. F.: Graphcut textures: image and video synthesis using graph cuts. *ACM Trans. Graphics (Proc. SIGGRAPH) 22*, 3 (2003), 277–286. 4

[KZ04] KOLMOGOROV V., ZABIH R.: What energy functions can be minimized via graph cuts? *IEEE TPAMI 26*, 2 (2004), 147–159. 2, 5

[LLS*09] LING C.-H., LIN C.-W., SU C.-W., LIAO H.-Y. M., CHEN Y.-S.: Video object inpainting using posture mapping. In *Proc. ICIP* (2009), pp. 2785–2788. 3

[MOG*06] MATSUSHITA Y., OFEK E., GE W., TANG X., SHUM H.-Y.: Full-frame video stabilization with motion inpainting. *IEEE TPAMI 28*, 7 (2006), 1150–1163. 3

[PKVP09] PRITCH Y., KAV-VENAKI E., PELEG S.: Shift-map image editing. In *Proc. ICCV* (2009), pp. 151–158. 2, 3, 4, 5, 8

[PPP11] PRITCH Y., POLEG Y., PELEG S.: Snap image composition. In *MIRAGE* (2011), Gagalowicz A., Philips W., (Eds.), vol. 6930 of *Lecture Notes in Computer Science*, Springer, pp. 181–191. 7

[PSB05] PATWARDHAN K. A., SAPIRO G., BERTALMIO M.: Video inpainting of occluding and occluded objects. In *Proc. ICIP* (2005), pp. 69–72. 3

[PSB07] PATWARDHAN K., SAPIRO G., BERTALMIO M.: Video inpainting under constrained camera motion. *IEEE TIP 16*, 2 (February 2007), 545–553. 3

[Rob99] ROBERTS A.: Fast and accurate multigrid solution of Poissons equation using diagonally oriented grids. *Numerical Analysis* (July 1999). 9

[SLCF06] SHEN Y., LU F., CAO X., FOROOSH H.: Video completion for perspective camera under constrained motion. In *Proc. ICIP* (2006), vol. 3, pp. 63–66. 2

[SMTK06] SHIRATORI T., MATSUSHITA Y., TANG X., KANG S. B.: Video completion by motion field transfer. In *Proc. IEEE CVPR* (2006), pp. 411–418. 3

[STH09] SHIH T. K., TANG N. C., HWANG J.-N.: Exemplar-based video inpainting without ghost shadow artifacts by maintaining temporal continuity. *IEEE Trans. Circuits Syst. Video Techn. 19*, 3 (2009), 347–360. 3

[STTHY08] SHIH T. K., TAN N. C., TSAI J. C., H.-Y Z.: Video falsifying by motion interpolation and inpainting. In *Proc. IEEE CVPR* (2008), pp. 1–8. 3

[VCZ09] VENKATESH M. V., CHEUNG S. S., ZHAO J.: Efficient object-based video inpainting. *Pattern Recognition Letters 30*, 2 (2009), 168–179. 3

[Wan95] WANDELL B. A.: *Foundations of Vision*. Sinauer Associates, Inc., 1995. 2, 6

[WSI07] WEXLER Y., SHECHTMAN E., IRANI M.: Space-time completion of video. *IEEE TPAMI 29*, 3 (2007), 463–476. 2, 3, 4, 5, 7, 8