**CGI2013 manuscript No.**
(will be inserted by the editor)

# Video Retargeting Combining Warping and Summarizing Optimization

**Abstract** We construct a unified interactive video retargeting system for video summarization, completion and reshuffling. Our system combines the advantages of both video warping and summarizing processing. We first warping the video to present initial editing results, then refine the results using patch-based summarizing optimization, which mainly eliminates possible distortion produced in the warping step. We develop a Mean Value Coordinate (MVC) warping method due to its simplicity and efficiency used in the initialization. For refining processing, the summarization optimization is built on a 3D bidirectional similarity measure between the original and edited video, to preserve the coherence and completeness of the final editing result. We further improve the quality of summarization by applying color histogram matching during the optimization, and accelerate the summarization optimization using by a constrained 3D Patch-Match algorithm. Experiment results show that the proposed video retargeting system effectively supports video summarization, completion, reshuffling while avoiding issues like texture broken, video jittering, detail losing.

**Keywords** Video summarization · video retargeting · video completion · bidirectional similarity · texture synthesis

## 1 Introduction

Motion picture and video are traditionally produced for specific displaying platforms such as cinema or TV. In recent years, however, we witness an increasing demand for displaying video content on devices with considerably differing display formats. Video retargeting, which focuses on presenting content-aware modification of the video for a comfortable viewing experience, have been intensively investigated, see [18] for a survey. The main objectives of the existing media retargeting methods can be described as following three aspects: preserving the important content of the original media; limiting visual artifacts in the resulting media; and preserving internal structures of the original media. Although the existing methods produce very promising results, however, there is still plenty of room to develop more sophisticated retargeting algorithms for improving both results and efficiency, and many challenges are left for video retargeting and summarization.

Several video retargeting approaches have been proposed. Most previous methods work by extending per-frame image-based techniques with some temporal considerations. Cropping methods [4, 13, 3] may produce virtual camera motions and artificial scene cuts, and important objects might be discarded. Although constraining temporally-adjacent pixels are used in [30, 16, 32, 12], due to camera and dynamic motion, the objects may deform inconsistently between frames, which will induce waving artifacts. Wang et al. [25] addressed this temporal coherence problem by explicit detection of camera and object motions, which only alleviates the waving artifacts. As the spatial limitation affects most video retargeting methods, by combining warping with temporally-based cropping, Wang et al. [26] partially overcame this spatial limitation. However, this approach still introduce such artifacts as virtual camera motions, when salient objects exhibit drastic motion, the retargeted results will not be natural.

F. Author
first adress

S. Author
second address

As an alternative, Simakov et al. [22] provided a visual data summarizing method using bidirectional similarity measure, which comprises two terms (completeness and coherence terms) between pairs of visual data (images or videos) to quantitatively capture these two requirements. Barnes et al. [2] later proposed a randomized algorithm for quickly finding approximate nearest neighbor matches to accelerate the summarization computing. Different from other media retargeting methods, Simakov et al. [22] exploited repetitiveness or redundancy of visual data in the summarization process, and can produce visually coherent small-sized summaries which are difficult to obtain with currently existing methods. Furthermore, this method is a powerful and versatile video editing tools, can be used for image (video) retargeting (summarizing), collages, reshuffling, and automatic cropping. However, effective initialization and computational efficiency are bottlenecks to make this method more practical.

In this paper, we present a powerful video retargeting approach combining the advantages of both the content-aware warping and patch-based summarization. Our method optimally combine them together to minimize visual artifacts in the summarized media. We summarize the video data using a 3D bidirectional similarity measure, which is derived from the bidirectional similarity measure [22]. As the summarizing processing is patch-based optimization operator, thus effective initialization and computational efficiency are critical for developing a practicable summarizing tool. To address these problems, we use a content-aware 3D mean-value coordinate (MVC) video warping approach, the warped results is "close enough" to the solution, works as the initialization value of the video summarizing system. Furthermore, using the correspondence between the source video and the warped video produced during MVC warping, we proposed a 3D approximately nearest neighbor search method to accelerate the video summarization processing. Guided by the mesh-correspondence as prior information, the nearest neighbor search is much faster and more accurate. The proposed approach can be used to address a variety of other problems, including video reshuffling and video object removal.

This paper makes the following three main contributions:

- We combine the advantages of both warping and summarization methods, where the warping method gives an initial edit result which is then refined by summarization method.
- We present an efficient video summarizing tool, which is based on 3D bidirectional similarity measure.
- We propose a MVC mesh constrained Patch-Match method for nearest neighbor searching between videos.

## 2 Related work

**Image retargeting**: Most content aware methods attempt to take advantages from the detection of pixel prominence. They either discard or distort the homogeneous regions in order to absorb the resulting distortion when changing the resolution of an image. These methods can be roughly classified into cropping methods [14, 23, 20], seam carving methods [1, 16], warping methods [8, 30, 32, 24], and multi-operators techniques [17] integrating crop and carve seams. As an alternative, Pritch et al. [15] presented a Shift-Map technique that allows removing a band region at a time, instead of a pixel-wise seam used in the seam carving methods, enabling the removal of entire objects. Recently, Rubinstein et al. [18] made comprehensive perceptual study and analysis on existing image retargeting approaches, and presented a methodological approach for evaluating seven retargeting methods.

**Video retargeting**: Video retargeting using cropping [4, 13, 3] may produce controlled virtual camera motions. Virtual camera motion may be quite large when processing video with dramatically temporal dynamics, and important objects might be discarded completely. Image resizing methods were extended to video by constraining temporally-adjacent pixels into the retargeting optimization [30, 16, 32]. Due to camera and dynamic motion, temporally-adjacent pixels do not necessarily contain corresponding objects, so that objects may deform inconsistently between frames, resulting in waving artifacts. Wang et al. [25] partially addressed this temporal coherence problem by explicit detection of camera and object motions. The spatial limitation affects most video resizing methods[30, 16, 32, 25]: if salient objects cover the entire frame space, their temporally consistent resizing degenerates into linear scaling. By combining warping with temporally-based cropping, Wang et al. [26] utilized degrees of freedom in the time dimension to overcome this spatial limitation. Although this approach produced pleasing results, however, it will introduce such artifacts as virtual camera motions. Additionally, this approach still can not work well when the salient object cover a large part of the video, and can not produce visually coherent small sized summaries. Later, they used per-frame optimization [27]to further improve their work.

**Data summarization**: Simakov et al. [22] considered the problem of image and video retargeting as a maximization of bidirectional similarity between small patches in the original and output images, and

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

Fig. 1: Video retargeting and comparisons. (a) Source video, (b) uniform scaling, (c) result of Want et al. [26], (d) Our summarization result.

this approach can be used to address a variety of other problems, including automatic cropping, completion and synthesis of visual data, photo reshuffling, object removal. A similar objective function and optimization algorithm was independently proposed by Wei et al. [28] as a method to create texture summaries for faster texture synthesis. Unfortunately, the approach of Simakov et al. is slow to minimize a path-based optimization. Barnes et al. [2] proposed a randomized algorithm for fast nearest neighbor matching, which accelerates the summarization computing to obtain interactive speeds for image editing with moderate size. More recently, using the high-level symmetry semantic, Wu et al. [31] presented a image resizing method by symmetry summarization, which removes or replicates the repetitive elements/regions in a more semantic fashion, and hence can better preserve the image symmetry structure.

**Mean-Value Coordinates**: Floater [6] introduced the Mean-Value Coordinates (MVC) which are motivated by the Mean-Value Theorem for harmonic functions. These coordinates approximate a harmonic-like solution to the boundary interpolation problem. They are well defined over the entire plane for arbitrary planar polygons without self-intersections, smooth, and invariant under similarity transformations [9]. MVC coordinates have also been extended to 3D polyhedra and used for space deformation [10, 7]. More recently, Farbman et al. [5] proposed a mean-value coordinate approach to accelerate seamless Poisson image and video cloning, where rather than solving a large linear system to perform Poisson interpolation, the value of the interpolant at each interior pixel is given by a weighted combination of values along the boundary. In this work, we explore the novel use of MVC as a alternative for solving video warping (retargeting).

## 3 Initialization by MVC Warping

The basic mechanism behind video warping methods is non-uniformly deforming the grids defined on video data, where the grids on important regions remain unchanged while grids on less salient regions are seriously distorted. This leads to visible distortion artifacts when background region is complex. For example, in the warping result produced by Wang et al. [26] (Figure 1 (c)), though the man in the foreground is preserved well, the doors in the background are squeezed too much. The method in this paper tries to solve this problem. We use the warping result as an initial solution, then we refine it using summarization method based on proposed 3D bidirectional similarity measure. Figure 1 (d) shows our final result.

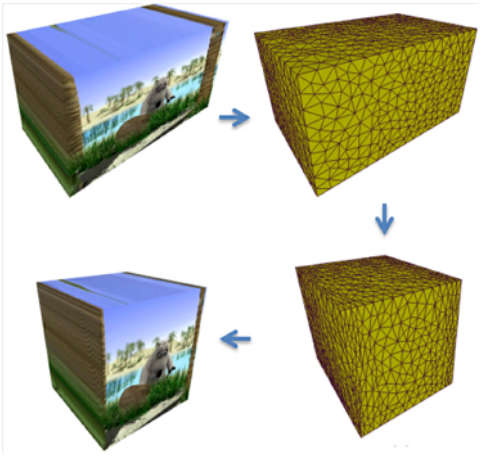


Fig. 2: The top left is source video, on which a tetrahedral mesh is built (top right). The source mesh is then warped into its half size (the bottom right). At last, the output video is received by interpolation.

Instead of using grids in each frame [26], we employ the TetGen [21], a tetrahedral mesh generator, to generate a tetrahedral mesh for video data (Figure 2 (b)). Then based on the tetrahedral mesh, we develop a content-aware video warping method based on 3D mean value coordinate (MVC), due to its simplicity and efficiency. Warping is transformed into a parametrization problem of finding deformed mesh which represent-

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

Fig. 3: MVC warping and video summarization incorporating MVC warping.(a) Source video, (b) uniform scaling, (c) MVC warping, (d) Our summarization results produced by correcting the MVC warping result.

ing the warped video, as illustrated in Figure 2 (c). We represent video using tetrahedron mesh $M = \{V, E, Q\}$, where $V = \{v_1, v_2, ..., v_n\} \in R^3$ is the set of vertex positions, $E$ and $Q$ denote the edges and tetrahedron faces, respectively. The new deformed vertex positions are denoted by $V = \{v'_1, v'_2, ..., v'_n\} \in R^3$. The connectivity is unalterable during the warping process.

To emphasize salient objects in video, we calculate saliency map (including the object motion)for the source video using the similar methods in [25, 26]. Each vertex $v_i$ of mesh $M$ is then associated with a saliency value $\omega_i$ which is the average of surrounding pixels. The salient objects will have large $\omega$ value. The tetrahedral meshes with less salient vertices will be distorted, the salient objects should be preserved.

For each internal vertex $v_i \in V$, we define the mean-value coordinates $\lambda_{i,j}$ [7, 9] with respect to its one-ring neighbors $v_{i,1}, ..., v_{i,d_i}$, where $v_{i,1}, ..., v_{i,d_i}$ is a closed region $\Omega$, and $d_i$ is the number of vertices on $\Omega$. Based on the properties of mean-value coordinates [7, 9], $v_i - \sum_{j=1}^{d_i} \lambda_{i,j} v_{i,j} = 0$. When source mesh $S$ is deformed to target mesh $T$, the salient objects should be shape-preserved and the content deformation of the video should be smooth. To achieve this goal, for the corresponding vertex $v'_{i,j}$ of each $v'_i$ in $T$, the distortion energy expression $(v'_i - \sum_{j=1}^{d_i} \lambda_{i,j} v'_{i,j})^2$ should be as least as possible for the salient objects. That is, each vertex $v'_i$ should keep the same mean-value coordinates $\lambda_{i,j}$ respect to its neighbors $v'_{i,1}, ..., v'_{i,d_i}$ as vertex $v_i$ does. We define the total distortion energy by summing up the individual vertex energy term for each internal vertex $v'_1, ..., v'_n$ and adding the saliency weight $\omega_i$, such that more distortion would be allowed in areas of less significance:

$$F_S = \sum_{i=1}^{n} \omega_i (v'_i - \sum_{j=1}^{d_i} \lambda_{i,j} v'_{i,j})^2 \quad (1)$$

Apart from the shape-preserving term $F_S$ which preserve the salient objects, we further use a common regularization term $F_R$ to smooth the deformation difference between adjacent tetrahedrons $A_i$ and $A_j$, which is defined as $F_R = \sum_i u_i \sum_{j \in N(i)} \|A_i - A_j\|_F^2$. The degree of penalization is controlled by weights $u_i$, which are computed as content saliency: regions with high saliency deserve high weights to prevent serious distortion.

The MVC warping function consisting of the shape-preserving deformation term and regularization term is defined as following:

$$F = F(v'_1, v'_2, ..., v'_n) = F_S + \beta \cdot F_R \quad (2)$$

subject to video boundary constraints. The weight $\beta$ is used to balance the two energy terms. We find that $\beta = 2$ is effective for most cases. The energy function can be minimized using an iterative solver such as reconditioned conjugate gradients (PCG) [19]. Our video retargeting operation can be performed in $X$, $Y$, and $Z$ directions simultaneously, since the coordinates of the mesh vertices are not coupled in energy function. Finally, we can receive the final video warping results by interpolating the vertex value of the tetrahedrons in the deformed mesh $T$ using method [10]. Figure 2 (d) shows the 3D results generated using our MVC warping. In Figure 3 we show one frame of Figure 2, the character is preserved well, but the background regions near left and right boundaries are seriously distorted.

## 4 The 3D Bidirectional Similarity Measure

Simakov et al. [22] provided a summarizing method using bidirectional similarity, and claimed that a good summarization result must contain as much as possible information from the input data and should introduce as few as possible new artifacts that were not in the input data. A bidirectional distance measure with two terms (completeness and coherence terms) between pairs of data are provided to quantitatively capture

these two requirements.

$$d(S,T) = \overbrace{\frac{1}{N_S}\sum_{P\subset S}\min_{Q\subset T} D(P,Q)}^{D_{complete}(S,T)} + \overbrace{\frac{1}{N_T}\sum_{Q\subset T}\min_{P\subset S} D(Q,P)}^{D_{cohere}(S,T)} \tag{3}$$

where $S$ is the source visual data, $T$ is the target data, $D$ is the SSD(sum of squared differences) of patch ($P$ and $Q$) pixel value in color space. $N_S$ and $N_T$ are the number of the patches used in $S$ and $T$, respectively.

It has been shown that this method can be used for image retargeting (summarizing), collages, reshuffling and so on. However, when processing video, it is important to keep the temporal consistency of the video content. We develop a 3D bidirectional similarity measure as the video summarization tool in our system:

$$d(S,T) = \frac{\overbrace{\sum_{P\subset S} W_P D(P,\tilde{Q})}^{D_{complete}(S,T)}}{\sum_{P\subset S} W_P} + \frac{\overbrace{\sum_{Q\subset T} W_{\tilde{P}} D(Q,\tilde{P})}^{D_{cohere}(S,T)}}{\sum_{Q\subset T} W_{\tilde{P}}} + \overbrace{\frac{1}{N_T}\sum_{Q\subset T} D(Q,Q')}^{D_{consistent}(S,T)} \tag{4}$$

Compared with (3), we add in Equation (4) a consistency term, which ensures that the summarization result is smooth and temporally consistent. Here, $Q$ is still a 3D patch in the output video, $Q'$ is a 3D patch in the initial result $I$ generated using MVC warping, and $Q'$ have the same location of $Q$. $\tilde{P} = \arg\min_{P\subset S} D(Q,P)$ and $\tilde{Q} = \arg\min_{Q\subset T} D(P,Q)$. $W_P$ and $W_{\tilde{P}}$ are two adaptive weights described in next section.

Though the warping result has distortion artifacts, it is consistent in the temporal space without jittering. With the consistency term $D_{consistent}$, we use the inherent consistency in warping result to help keep consistency in summarization processing. In Figure 5, we give the comparison results with [22]. The method of [22] produces blurring results due to the weighted average of inaccurate patch value used in the optimization computing. Our 3D bidirectional similarity measure produces more consistent result.

In order to reduce computation time, inspired by [11] (where the authors synthesize solid texture from exemplar image), we only measure the differences on the three slices orthogonal to the main axes of the video volume, instead of measuring the whole 3D patches, as illustrated in Figure 4. The distance form 3D patch $P$ to $Q$ is defined as:

$$D(P,Q) = \sum_{i\in\{x,y,z\}} \|P_{v,i} - Q_{v,i}\|^r. \tag{5}$$

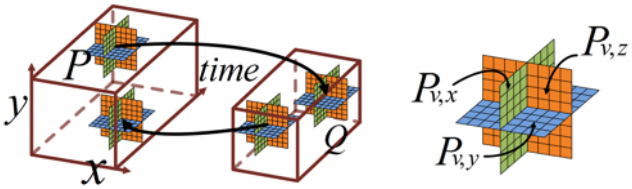
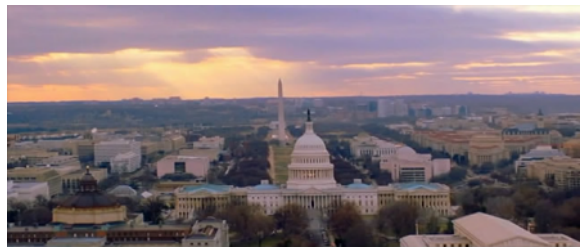
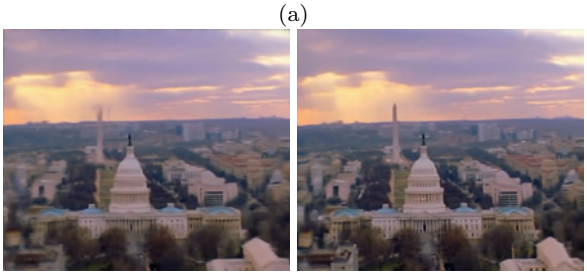



Fig. 4: 3D patch similarity measure using the differences on the three slices orthogonal to the main axes of the video volume.



(a)



(b) (c)

Fig. 5: Summarization comparisons (a) Source video, (b) results of [22], (c) Our result of 3D bidirectional similarity.

Here $\nu$ refers to a single voxel, which is the center of 3D patch $P$, and $P_{\nu,x}$, $P_{\nu,y}$ and $P_{\nu,z}$ are the vectorized neighborhoods of $v$ in the slices orthogonal to the x, y, and z axis, respectively. $Q_{\nu,x}$ , $Q_{\nu,y}$ and $Q_{\nu,z}$ are defined in the similar way. The exponent $r = 0.8$ causes the optimization to be more robust against outliers. Compared with solid 3D patch similarity measure, the proposed method is much faster, in addition, the blurring artifacts can be alleviated.

### 4.1 Optimization Update Rule

We obtain an output video by minimizing equation (4):

$$T_{output} = arg\min_{T} d(S,T). \tag{6}$$

We use EM algorithm to solve Equation (6). Let $p \in S$ be a pixel in $S$, and let $S(p)$ be its color. In the $(l+1)^{th}$ iteration, for each pixel $q$ in $T^{l+1}$, its value is the average of three kinds of pixels in the input video

$S$, and the initial MVC warping result $I$, respectively. We set $T^0$ as $I$:

1. **In the coherence term**: for all patches $Q_1, ..., Q_m$ containing $q$, there will be corresponding (most similar) patches $\tilde{P}_1, ..., \tilde{P}_m$ in $S$. Thus, the first kind of pixels are $\tilde{p}_1, ..., \tilde{p}_m$ in $\tilde{P}_1, ..., \tilde{P}_m$ corresponding to the location of pixel $q$ within $Q_1, ..., Q_m$. Here, $m$ is the number of pixels in patch.
2. **In the complete term**: suppose there are $n$ patches $P_1, ..., P_n$ in $S$, whose nearest patches in $T^l$ are $\tilde{Q}_1, ..., \tilde{Q}_n$, and $\tilde{Q}_1, ..., \tilde{Q}_n$ just contain pixel $q$. The second kind of pixels are $p_j$ in $P_j$ corresponding to the location of pixels $q$ within $\tilde{Q}_j$. Note that $n$ may be zero if no patch in $S$ points to a patch containing $q \in T^l$ as its most similar patch.
3. **In the consistent term**: similar to the coherence term, for each patch $Q_k$ containing $q$, there will be a corresponding patch $Q'_k$ in the initial MVC warping video $I$. Thus, the third kind of pixels are $q'_k$ in $Q'_k$ corresponding to the location of $q$ within $Q_k$. The number of such kind of pixels is also $m$.

According to the above analysis, we compute the new value for pixels $q$ in output video $T^{l+1}$ as:

$$T^{l+1}(q) = \left( \frac{\sum_{j=1}^{n} W_{P_j} S(p_j)}{\sum_{j=1}^{n} W_{P_j}} + \frac{\sum_{i=1}^{m} W_{\tilde{P}_i} S(\tilde{p}_i)}{\sum_{i=1}^{m} W_{\tilde{P}_i}} + \frac{1}{N_T} \sum_{k=1}^{m} I(q'_k) \right) / \left( \sum_{j=1}^{n} W_{P_j} + \sum_{i=1}^{m} W_{\tilde{P}_i} + \frac{m}{N_T} \right) \tag{7}$$

Figure 1 (d), Figure 3 (d) and Figure 5 (c) shows results of our summarization method. Please refer to supplemental video for more apparent comparison.

### 4.2 Adaptive Weights Using Histogram Matching

Iterative EM optimization may lead to blurring in the output video. Inspired by [11], we consider assigning adaptive weights $W_P$ and $W_{\tilde{P}}$ for pixels joining in the averaging process, which gives more weights to important pixels. This way ensures that the histogram of output is similar to that of source video, which alleviates the risk of falling into the local minima during the optimization processing and improves the results.

In each M-step, we build a histogram with 32 bins for each patch in both source video and output video. Let $H_{P,j}$ be the histogram of $j^{th}$ ($j \in 1, 2, 3$ if using RGB color space) channel of patch $P$. $H(b)$ be the value of $b^{th}$ bin of the histogram, the difference of two histograms is $\Im(H_P, H_Q) = \sum_{j=1}^{3} \sum_{b=1}^{32} (H_{P,j}(b) - H_{Q,j}(b))$. In the $(l+1)^{th}$ iteration, the new weights $W_P^{l+1}$ for patch $P$ after histogram matching is defined as:

$$W_P^{l+1} = W_P^l / \Im(H_P, H_{\tilde{Q}}) \tag{8}$$

where, $W_P^l$ is the weight of patch $P$ in the last iteration, $\tilde{Q}$ is the most nearest patch corresponding to $P$. At the very beginning, $W_P^0$ is the average saliency value computed in the warping step.

Figure 6 gives a comparison before and after histogram matching, which shows that the histogram matching alleviates the blurring artifacts greatly in the summarization results.

## 5 Accelerating By Mesh Constrained 3D Patch-Match

Barnes et al. [2] proposed a randomized approximately nearest neighbor matching algorithm for structure image editing. The algorithm is mainly based on two key observations: Firstly, the natural coherence in the structural image allows propagating such good matches quickly to surrounding areas; secondly, good patch matches can be found via random sampling. Barnes et al. [2] applied three main components, initialization, propagation and random search to implement the above two observations. In this paper, we extend this approach on video summarization. However, applying this method directly on 3D patch matches is extremely slow, furthermore, as video summarization is more complicated than image summarization. To address above problem, we incorporate the MVC warping results into the summarization process, and using the warping mesh to constrain the Propagation stage in the Patch-Match method.

**Mesh Constrained Propagation**: As the 2D Patch-Match, we attempt to improve patch pairs $f(x, y, z)$ using the known offsets of $f(x-1, y, z)$, $f(x, y-1, z)$, $f(x, y, z-1)$. Let $D(v)$ denote the patch pair distance between the patch at $(x, y, z)$ in $S$ and patch $(x, y, z)+v$ in $T$, we take the new value for $f(x, y, z)$ to be the $\arg\min$ of $\{D(f(x, y, z)), D(f(x-1, y, z)), D(f(x, y-1, z)), D(f(x, y, z-1))\}$. The basic idea is that if $(x, y, z)$ has a correct mapping and is in a coherent region $R$, then all of patch pairs in $R$, back, below and right of $(x, y, z)$ will be filled with the correct mapping. Moreover, on even iterations, to improve $f(x, y, z)$, we propagate information up, front and left by examining patch pairs in reverse scan order, using $f(x+1, y, z)$, $f(x, y+1, z)$ and $f(x, y, z+1)$ as our candidate patch pairs.

Although the above propagation works well, but it is very slow for a moderate video. We constrain the propagation domain to improve speed. Suppose a tetrahedron

(a) (b) (c)

Fig. 6: Video retargeting incorporating histogram matching and comparisons. (a) source video, (b) result without histogram matching, (c) result with histogram matching.

(mesh) $A$ in the source video is deformed into tetrahedron $A'$ in the retargeted video, the nearest patch of pixel $p \in A$ should be constrained in $A'$ or its neighboring tetrahedrons. Similarly, the nearest patch of pixel $q \in A'$ should be constrained in $A$ or its neighborhoods. Thus, in our method, we not only perform propagation using the image structure, but also constrain the nearest patch in a specified regions. Using this strategy, we not only accelerate nearest patch matching, but also avoid the artifacts such as undesirable local minima artifacts. Note that in video summarization, it is important to keep the temporal consistency. In nearest neighbor searching, we have to give more restrict for the neighbor search in temporal direction, and penalize the violation of temporal relations between the objects.

## 6 Results and discussion

Our system targets towards high level video editing, and can generate video retargeting and summarization results both automatically and interactively. We implemented our algorithm on a computer with 2 Xeon(R) CPUs at 2.27GHz and 4GB RAM. Our MVC warping typically takes less than 2 seconds to process an video streaming with size of $640 \times 360 \times 120$, while the computation time also depends on the resolution of the tetrahedron mesh, we usually build a mesh with resolution about $50 \times 30 \times 20$ for aforementioned video. The main time-consuming step is the nearest neighbor matching in our summarization system. It usually takes between 8 and 15 minutes to summarize aforementioned size of video to a target video with size of $320 \times 360 \times 120$ on CPU, similarly, the computation time also depends on the size of the target video, and the number of the iteration.

In this section, besides of the video retargeting, we show that the proposed approach can be used to address a variety of other problems, including video summarization, completion, reshuffling. We also present the comparison results with the state-of-the-art video retargeting and summarizations methods.

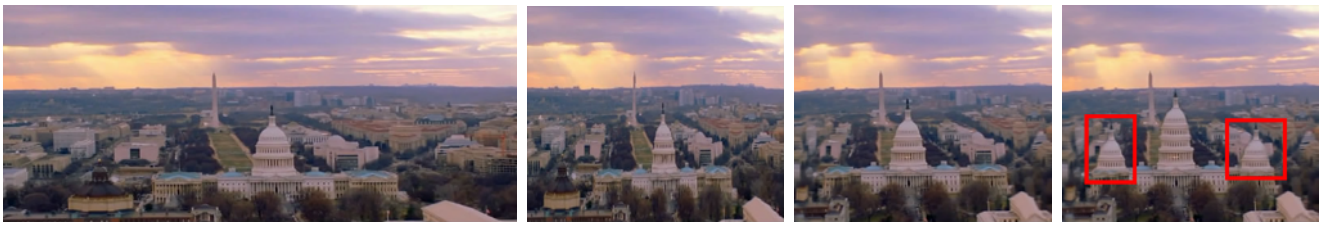
**Video retargeting**: Our system can achieve high-quality video retargeting to arbitrary aspect ratios for complex videos containing diverse camera and dynamic motions. Most previous content-aware retargeting methods concentrated on spatial considerations, attempting to preserve the shape of salient objects in each frame by removing or distorting homogeneous less important content, however, these methods may cause waving and squeezing artifacts due to fundamentally limited space used for sacrificing. By employing motion information, we first warp the video, which distributes distortion in both spatial and temporal dimensions, then based on the warping results, we correct the warping results using the summarizing techniques. Our method can retarget challenging videos with complex motions, numerous prominent objects, even the video with foreground and background regions heavily correlated.

Our method compares favorably with state-of-the-art retargeting systems. In Figure 1, Figure 5, Figure 7, and Figure 8, we compare with several other competing methods [22, 2, 26]. The bidirectional similarity measure [22] can be extended to video summarization. However, due to the weighted average of inaccurate patch value used in the optimization computing, some content in the video is blurred, as illustrated in Figure 5. The method [22] can be accelerated using the randomized patch match method [2]. As the nearest patch match is randomized, the search results may not be accurate, and the method [2] may fall into local minima, which may leads to undesirable results such as distortion (Figure 7). Compared with our method, more iterations and computation time are required to receive the final stable results using the gradual resizing [22].

Wang et al. [26] utilized freedom degree in the time dimension to overcome spatial limitation in video retargeting, and received pleasing results for most input videos. However, by combining warping with temporally-based cropping, this approach will introduce such arti-

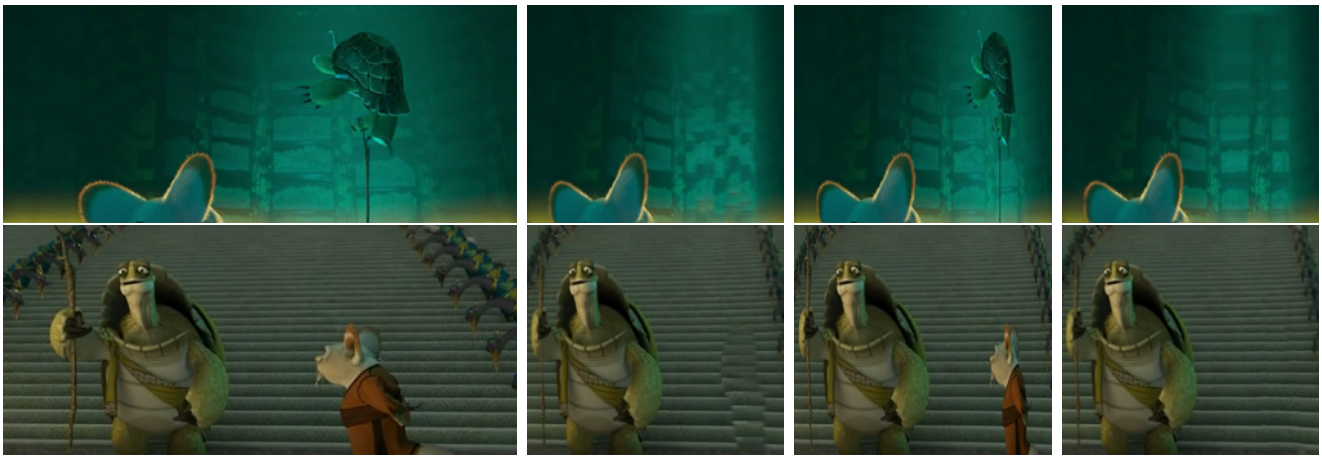
(a) (b) (c) (d)

Fig. 7: Video retargeting comparisons, (a) source video, (b) result of [22] accelerated using [2], (c) video retargeting of [26], (d) result of our method. Note that all animals are preserved.



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

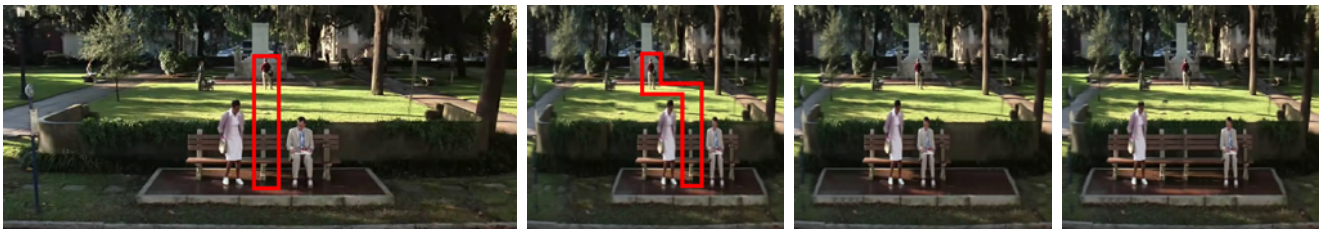
Fig. 8: Video retargeting and reshuffling, (a) source video, (b) video retargeting of [26], (c) results of our method, (d) video reshuffling results using our method.



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

Fig. 9: video object removal, (a) source video, (b)object removal using [22], (c) MVC warping, (d) object removal using our approach.

facts as virtual camera motions, when salient objects perform drastic motion, the retargeted results will not be natural, as shown in Figure 7. Additionally, the method [26] is essentially a warping method. As shown in Figure 1, though the man in the foreground is preserved well, the doors in the background are squeezed too much (Figure 1 (c)). However, our method can be considered as a patch-based texture optimization, it produces much better results, as illustrated in Figure 1(d). Finally, this approach still may not work well when the salient object covers a large part of the video, or when there are too many important objects in the video. In these cases, when retargeting the source video into a small target video, their temporally consistent resizing degenerates into linear scaling, or many objects have to be cropped out or be squeezed with heavy distortion, as illustrated in Figure 8.



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

Fig. 10: Video reshuffling, (a)source video, (b) object is shifted, (c) the video is summarized, and some parts of chair is removed, (d) the video is summarized, and some parts of chair is cloned.

**Video Completion:** Video completion of large missing regions is a challenging task. Even the advanced global optimization method [29] can still produce inconsistencies where structured content needs to be completed. In many cases the boundaries of the missing region provide few or no constraints for a plausible completion. Importance weights $\omega$ can also help to remove undesired objects from the target video in the summarization process. Combining importance weights with the constraint of nearest neighbor search regions, our video summarization system can be used for video completion and object removal, as shown in Figure 9.

We have compared performance and quality with competing methods [22, 2] in video completion. As shown in Figure 9, the tortoise in the source video is successfully removed in the summarized video, and the structures are well completed, which is consistent with the boundary content. Using the randomized correspondence algorithm [2], we observe that there are "ghosting" artifacts left in the summarized video. These artifacts happen because of the randomized initialization for energy optimization and the randomized nearest neighbor search. Note that for video completion, it is important to keep the temporal consistency, thus, in nearest neighbor search, we have to give more restrict and larger weight for the neighbor search in temporal direction, and penalize the violation of temporal relations between the objects.

**Video reshuffling**: Based on the mesh corresponding information built during the MVC warping, incorporating the constrained randomized correspondence and user interactivity, our system can effectively manipulate the video reshuffling. Our video reshuffling works as follows, we first warp the source video into the target video using the MVC method. Then in the target video, we drag the user specified object to the destination location. Based on the mesh corresponding information built during the MVC warping, we constrain the regions for nearest neighbor search to synthesize the specified object. The hole left by the specified object in the original region can be completed using our system provided in the above section. With appropriate manual intervention, our system can remove, swap, copy, stretch, and zoom the user specified objects in the video, while keeping the video reshuffling results spatially consistent and temporally coherent.

As illustrated in Figure 8 and Figure 10, we give several reshuffling results. The object is moved, enlarged, or parts of the object are cloned. Our system is more convenient for user interaction than [22, 2]. The user constraints described in MVC warping method can succeed in preserving lines and regions, and define a satisfied initialization for bidirectional similarity measure. Combining with powerful and versatile editing tool of patch-based optimization, our system can gradually rearrange the video content to align with these constrained regions, and reshuffle objects automatically.

**Limitation**: Although the proposed method greatly speed the nearest neighbor search process, however, we still can not obtain interactive video editing summarization, we would like to further accelerate the nearest neighbor search, one possible approach is to accelerate the speed on the graphics hardware-GPU. Similar with all video retargeting and summarization approaches, our algorithm does have some failure cases, for example, extreme edits to an video can sometimes produce unsatisfied results, where our MVC warping can not produce a plausible results, and the summarizing step can not correct the results due to extreme edits.

## 7 Conclusion and future work

We proposed an efficient video data summarization system combining content-ware warping and patch-based optimization. Our method combines both the advantages of video warping and patch-based video optimization, the visually important regions can be well preserved, while the non important regions can be efficiently removed or squeezed with respect to the desired scaling factors. Our system can produce spatially

and temporally coherent video summarization results for complex video containing dynamic motions.

In the future, we will construct a more flexible adaptive tetrahedralization method for the video volume, the tetrahedralization method should be more edge and saliency aware, which may result in more accurate warping results. In addition, we would like to work on the video synopsis which retarget the video both the spatial and temporal direction.

## References


1. Avidan S, Shamir A (2007) Seam carving for content-aware image resizing. ACM Transactions on Graphics 26(3):10
2. Barnes C, Shechtman E, Finkelstein A, Goldman D (2009) PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics (TOG) 28(3):1–11
3. Deselaers T, Dreuw P, Ney H (2008) Pan, zoom, scanTime-coherent, trained automatic video cropping. In: CVPR 2008, IEEE, pp 1–8
4. Fan X, Xie X, Zhou H, Ma W (2003) Looking into video frames on small displays. In: Proceedings of the eleventh ACM international conference on Multimedia, ACM, p 250
5. Farbman Z, Hoffer G, Lipman Y, Cohen-Or D, Lischinski D (2009) Coordinates for instant image cloning. In: ACM SIGGRAPH 2009 papers, ACM, pp 1–9
6. Floater M (2003) Mean value coordinates. Computer Aided Geometric Design 20(1):19–27
7. Floater M, Kós G, Reimers M (2005) Mean value coordinates in 3D. Computer Aided Geometric Design 22(7):623–631
8. Gal R, Cohen-Or D (2006) Feature-aware texturing. Rendering Techniques 2006
9. Hormann K, Floater M (2006) Mean value coordinates for arbitrary planar polygons. ACM Transactions on Graphics (TOG) 25(4):1424–1441
10. Ju T, Schaefer S, Warren J (2005) Mean value coordinates for closed triangular meshes. In: ACM SIGGRAPH 2005 Papers, ACM, pp 561–566
11. Kopf J, Fu C, Cohen-Or D, Deussen O, Lischinski D, Wong T (2007) Solid texture synthesis from 2d exemplars. ACM transactions on graphics 26(3):2
12. Krahenbuhl P, Lang M, Hornung A, Gross M (2009) A system for retargeting of streaming video. ACM Transactions on Graphics (TOG) 28(5):1–10
13. Liu F, Gleicher M (2006) Video retargeting: automating pan and scan. In: Proceedings of the 14th annual ACM international conference on Multimedia, ACM, pp 241–250
14. Liu H, Xie X, Ma W, Zhang H (2003) Automatic browsing of large pictures on mobile devices. In: Proceedings of the eleventh ACM international conference on Multimedia, ACM, pp 148–155
15. Pritch Y, Kav-Venaki E, Peleg S (2010) Shift-map image editing. In: ICCV, IEEE, pp 151–158
16. Rubinstein M, Shamir A, Avidan S (2008) Improved seam carving for video retargeting. ACM Transactions on Graphics-TOG 27(3):16–16
17. Rubinstein M, Shamir A, Avidan S (2009) Multi-operator media retargeting. ACM Transactions on Graphics (TOG) 28(3):1–11
18. Rubinstein M, Gutierrez D, Sorkine O, Shamir A (2010) A comparative study of image retargeting. In: ACM Transactions on Graphics (TOG), ACM, vol 29, p 160
19. Saad Y, Saad Y (1996) Iterative methods for sparse linear systems, vol 620. PWS publishing company Boston
20. Santella A, Agrawala M, DeCarlo D, Salesin D, Cohen M (2006) Gaze-based interaction for semi-automatic photo cropping. In: Proceedings of the SIGCHI conference on Human Factors in computing systems, ACM, pp 771–780
21. Si H, TetGen A (2006) A quality tetrahedral mesh generator and three-dimensional delaunay triangulator. Weierstrass Institute for Applied Analysis and Stochastic, Berlin, Germany
22. Simakov D, Caspi Y, Shechtman E, Irani M (2008) Summarizing visual data using bidirectional similarity. In: CVPR 2008., IEEE, pp 1–8
23. Suh B, Ling H, Bederson B, Jacobs D (2003) Automatic thumbnail cropping and its effectiveness. In: ACM symposium on User interface software and technology, ACM, pp 95–104
24. Wang Y, Tai C, Sorkine O, Lee T (2008) Optimized scale-and-stretch for image resizing. ACM Trans Graph 27(5):118
25. Wang Y, Fu H, Sorkine O, Lee T, Seidel H (2009) Motion-aware temporal coherence for video resizing. ACM Transactions on Graphics (TOG) 28(5):1–10
26. Wang Y, Lin H, Sorkine O, Lee T (2010) Motion-based video retargeting with optimized crop-and-warp. In: ACM SIGGRAPH 2010 papers, ACM, pp 1–9
27. Wang Y, Hsiao J, Sorkine O, Lee T (2011) Scalable and coherent video resizing with per-frame optimization. In: ACM Transactions on Graphics (TOG), ACM, vol 30, p 88
28. Wei L, Zhou K, Han J, Guo B, Shum H (2008) Inverse texture synthesis
29. Wexler Y, Shechtman E, Irani M (2007) Space-time completion of video. IEEE transactions on pattern analysis and machine intelligence pp 463–476
30. Wolf L, Guttmann M, Cohen-Or D (2007) Non-homogeneous content-driven video-retargeting. ACM Transactions on Graphics
31. Wu H, Wang Y, Feng K, Wong T, Lee T, Heng P (2010) Resizing by symmetry-summarization. In: ACM Transactions on Graphics (TOG), ACM, vol 29, p 159
32. Zhang Y, Hu S, Martin R (2008) Shrinkability Maps for Content-Aware Video Resizing. In: Computer Graphics Forum, Wiley Online Library, vol 27, pp 1797–1804