Video Completion Based on Improved Belief Propagation

Yimin Yu, Duanqing Xu, Chun Chen, Lei Zhao
College of Computer Science,
Zhejiang University,
310027 Hangzhou, Zhejiang,
P.R.China

Abstract: - In this paper, a novel computer vision technique is introduced to fill the holes in a video sequence due to removal moving foreground objects. Contrary to existing greedy techniques, we regard the task as a global optimization problem, which can overcome the obstacle of visual inconsistencies and don't require user intervention. After the pixel extensions in the frame via our definitions, we can deal with the time information included in video sequence freely and well. By means of minimizing the energy defined on the 3d label graph, we could match the appropriate patches to fill the hole. Since the traditional standard BP will lead to efficient problem caused by a great deal of labels, we improve the standard belief propagation algorithm and apply a better algorithm called priority belief propagation. This algorithm use label pruning and priority based message scheduling method to reduce the computational cost. From the experiment results, we can discover our method can considerably speed up BP's convergence and the completion performance is impressive.

Key-Words: - video completion, global optimization, belief propagation, label pruning, message scheduling

1 Introduction

Video completion aims at filling the missing parts of frames in video sequences in a visually plausible way. It is solved by using information from other parts of the sequence to suggest suitable in-fill. Video completion has become viable in many video processing and recognition. It has many applications, in areas such as video editing and film postproduction including composition of graphic objects, removal of unnecessary objects, and insertion of virtual objects to a video sequence. Video completion plays a key role in a variety of research areas and has been emerged to be a high level understanding task of low level vision content. By means of this technique, most tasks that have been done by human intervention could be completed efficiently and perfectly now.

The technique of video completion advances rapidly owing to more achievement obtained in the area, while it is still a hotspot in computer vision domain. There are some obstacles to be overcome such as frame blurring and temporal consistency etc. Bhat et al. [1] extract a moving object from a video sequence by making a mosaic image of the scene. Sand and Telle [4] use two video sequences to match features frame by frame. Zhang et al. [8] separate foreground and background objects in multiple motion layers to extract a moving foreground object. The layers are non-overlapping and each of them is completed separately. Although the result is approving but it could not deal with the instance while transformation between frames. Sugaya and Kanatani [5] assume a feature point is an affine transformation of a 3D feature point and generate a trajectory vector by stacking 2D coordinates of tracked features from all frames. By affine space fitting, they remove outliers, extract and remove a moving object. Wexler et al. [7] addresses a problem of completing a moving object which is partly occluded in a video sequence. They use a stationary camera to obtain the video frame. They complete the occluded object areas by filling empty holes with space-time patches which have spatio-temporal consistency.

In this paper, a novel approach based on improved belief propagation algorithm is proposed to fill the blank region due to the removal of unwanted objects in the video sequences. At first we give some definition to extend the common planar processing concept. By means of adding dimensionality to contain time information, we can deal with video material easily. Then we propose a well-suited similarity measurement between patches centered at pixels. Contrary to greedy synthesis methods, we regard completion procedure as a discrete global optimization problem with a well defined objective function. Furthermore, we introduce a novel optimization algorithm called “Priority-BP”, which improve the standard BP algorithm. It carries 2 major improvements over standard belief propagation:

1). “label pruning”

2). “priority-based message scheduling”.

They bring a dramatic reduction in the overall computational cost of BP which would otherwise be intolerable due to the huge number of existing labels. With our method, no user intervention is required. We manage to avoid greedy patch assignments by maintaining (throughout its execution) many candidate source patches for each block of missing pixels.

The remaining part of this paper is organized as follows: First, in section 2, we present some definitions such as 3-dimension ST point and the patch distance between two ST patches. Section 3 proposes an idea of viewing video completion as a graph labeling problem. Next, we introduce the improved belief propagation algorithm in detail in section 4. Detail experimental results and effect comparison are shown in section 5.

2 Definition

Since video sequences contain dynamic objects correlated with time info, we could combine the space info with time info to an all-in-one framework. We adopt the space-time method [7] to illustrate the content. We extend the planar coordinate in each frame to a cubic coordinate by adding one dimension of time info. Thus, a pixel in the video sequences is denoted as $o(x, y, t)$ instead of $o(x, y)$. Given an input video sequence K, the hole is defined a target region R due to removal of unwanted objects, and a source region S which is subset of K-R.

Definition 1 ST coordinate is defined as:

2D planar coordinate: $o(x, y)$ ->3D ST coordinate: $o(x, y, t)$

Definition 2 the global visual coherence is defined as:

If each ST patch in one video sequence V_1 can be found in other video sequence V_2, then we call the two sequences are global visual coherence.

We hope the missing portion can be filled well, however, how to say the result sequence is better? So we use the definition above to solve the problem. Since we use the video portion of known region S to fill in the unknown region R, we should seek the maximum similarity between region R and S.

Definition 3 the attribute of ST pixel is defined as:

In a planar image, we usually take into account the color info of pixels, and it is effective for image completion. However, the video sequence has not only those but also the time info. Since the objects motion along the time is high sensitive to human eyes, we couldn't only take care of the common attribute. In fact, the object motion continuity is more important than the spatial color attribute.

Due to the reason above, an extended attribute is proposed as follows:

Set F for the video sequence containing the gray intensity and the ST point is $o(x, y, t)$. Then the spatial and temporal derivative is (F_x, F_y, F_t). Thus the horizontal motion can be characterized as: $u = \frac{F_t}{F_x}$, this attribute can be used to depict the instantaneous motion along the x direction. The vertical motion can be characterized as: $v = \frac{F_t}{F_y}$, and this attribute can be used to depict the instantaneous motion along the y direction. Finally, combining the new two attribute with original color attribute, we can get a 5-dimensional attribute for ST point: (r, g, b, u, v).

Definition 4 the distance and similarity is defined as:

To estimate the similarity between two ST patches, a straightforward way is to extend the common Sum of Squared Difference (SSD) of color attribute to a new one including time info. Based on the definition 3, we can realize it. Let two ST patches as P_s and P_t, we can get the distance:

$$d(P_s, P_r) = \sum_{(x,y,t)} \| P_s(x,y,t) - P_r(x,y,t) \|^2 \qquad (1)$$

$P_s(x, y, t)$ and $P_t(x, y, t)$ are the attribute vector (r, g, b, u, v) of each pixel on patch P_s and P_t respectively.

Definition 5 the similarity measurement is defined as:

A well-suited similarity measurement between patches is the heart of the algorithm that directly influences the final completion result. The exponential similarity of two patches can be derived from the distance defined in definition 4, and the detail expression can be shown as follows:

$$s(P_s, P_r) = e^{\frac{-d(P_s,P_r)}{2\sigma^2}} \qquad (2)$$

The σ reflects noise, it set to 5/255 graylevels.

3 Graph construction

Reviewing the image and video completion skills, we can discover that the example-based techniques using texture synthesize have been the most successful in dealing with the completion problem compared to statistical-based or PDE-based methods. In this section, we introduce the texture propagation algorithm. The problem we address is how to synthesize the unknown region by using sample patches in the known region. As noted above, by obtaining the motion parameters from the 5D vector (r, g, b, u, v), we can reduce the portion of the video to search for matching the unknown region.

According to [7], the completion result is satisfied if and only if the two criterions are satisfied:
1) For every ST point o all the ST patches $P_{o1}...P_{ok}$ which contain it agree on the color value at o.
2) All those $P_{o1}...P_{ok}$ appear in the dataset S (known region).

Let o be an unknown point and $o\in R$. Let C = $\{P_{o1} ,..., P_{ok}\}$ be all those ST patches containing o. Let D = $\{P_{r1} ,..., P_{rk}\}$ be the patches in S that are most similarity to $C = \{P_{o1} ,..., P_{ok}\}$ according to equation (2). Then our work is made clearly. We need only search appropriate patches in D to fill the unknown region R, and avoid much unnecessary computing to search the whole video sequence. It gives high efficiency. We set the patch size in the patches set D as 5×5×5, thus there are 125 different patches involve point o. We can transform this patch as larger patch whose center locates at o. So the new patch size is 9×9×9. Thus we reassign the patch set D as $D = \{D(1), D(2), \dots, D(n)\}$.

Unlike Wexler's [7] approach, we consider the region propagation in video sequence as a graph labeling problem, where each patch in the video sequence is assigned a unique label. We adopt the technique of [3] to construct a 3D graph $G = <V, A>$ on a video sequence. V is the node set of ST points in the unknown region, and we set the points $\{o_i\}_{i=1}^{m}$ as the node of V. The arc set A contains two kinds of arcs: intra-frame arcs A_I connecting adjacent nodes in one frame; inter-frame arcs A_O connecting adjacent nodes across adjacent frames. The intra-frame arcs A_I are constructed by connecting adjacent nodes in one frame, and the inter-frame arcs A_O are constructed by connecting each node to each other node in the adjacent frame $f_{t\pm1}$. The interval of sample patches in D is set to half of the planar patch size to guarantee sufficient overlaps.

For each unknown point o_i, we find a label $l_i \in \{1, 2, \cdots, n\}$ corresponding to one of the patches in D. Then we just fill the color info of selected patch $D(l_i)$ to the point o_i.

3.1 3D graph energy

The energy function on the graph G can be defined as follows:

$$E(l) = \sum_{i\in V} E_1(l_i) + \lambda_\alpha \sum_{(i,j)\in A_I} E_2(l_i,l_j) + \lambda_\beta \sum_{(i,j)\in A_O} E_3(l_i,l_j) \quad (3)$$

The E_1 is called likelihood energy and the E_2, E_3 is called prior energy. λ_α and λ_β are relative weights, and their default values are: $\lambda_\alpha = 22$, $\lambda_\beta = 11$. The optimal sample labels $L = \{l_i\}_{i=1}^{m}$ are obtained by minimizing the energy $E(l)$.
Likelihood energy:

$$E_1(l_i) = \psi_\alpha E_\alpha(l_i) + \psi_\beta E_\beta(l_i) \quad (4)$$

E_1 measures the conformity of the color of o_i to the color of known patch $D(l_i)$. The ψ_α and ψ_β are relative weights. The $E_\alpha(l_i)$ denotes the similarity between the sample patch and the node i in unknown region. Its function term is shown as:

$$E_\alpha(l_i) = \sum_{z\in q^2} s_i \left\| c(i+z) - c(l_i+z) \right\|^2 \quad (5)$$

The s_i is the robust parameter that weights the color contrast according to equation (2). $c(i + z)$ and $c(l_i + z)$ are color of point and the q is the area size of patch.

$E_\beta(l_i)$ constrains the synthesized patches on the boundary of unknown region R to match well with the known point in S. $E_\beta(l_i)$ is the sum of the normalized squared differences (SSD) calculated in out of the region R on boundary patches. The portion of $E_\beta(l_i)$ in region R is set to zero.

The prior energy E_2 and E_3 constrains the coherence between adjacent synthesized patches $D(l_i)$ and $D(l_i)$, where l_i and l_j are labels for adjacent nodes. E_2 measure color differences between two adjacent regions in the same frame, and the third term E_3 measures color differences between two adjacent patches in two adjacent frames, and embeds temporal coherence in the optimization process through intra-frame arcs A_I. In a similar fashion, the pair-wise potential $E_{ij}(l, l')$, due to placing patches l, l' over neighbors o_i, o_j, will measure how well these patches agree at the resulting region of overlap and will again be given by the equation similar to equation 5. In fact, E_2 and E_3 are penalty when adjacent nodes are assigned with different labels.

4 Improved belief propagation

4.1 Belief propagation

Belief propagation is an inference algorithm for graphical models containing continuous, non-Gaussian random variables. BP is a specific instance of a general class of methods that exist for approximate inference in Bayes Nets (variational methods) or Markov Random Field. Inference problems arise in statistical physics, computer vision, error-correcting coding theory, and AI. We explain the principles behind the belief propagation (BP) algorithm, which is an efficient way to solve inference problems based on passing local messages. Applying the technique of [6], we define a n-element vector M_{ij} that denotes the message sent from node o_i to node o_j. The vector M_{ij} indicates how likely node o_i

thinks that node o_j should be assigned label l_j. The detail process of the method has been presented as algorithm 1.

Algorithm 1: video completion with belief propagation

```
1  Initialize all messages $M_{ij}^0 = 0$ between any two
   adjacent nodes i and j in graph G.
2  Update all messages $M_{ij}^t$ iteratively from t = 1 to
   T:
   $M_{ij}^t = \min_{l_i}\{E_1(l_i) + E_2(l_i, l_j) + E_3(l_i, l_j) + \sum_{k \neq j, (k,i) \in A} M_{ki}^{t-1}\}$
3  The optimal label $\hat{l}_i$ for each node i:
   $\hat{l}_i = \arg\min_{l_i}\{E_1(l_i) + \sum_{(k,i) \in A} M_{ki}^T\}$
```

4.2 Node Priority

From algorithm 1, we set $B(l_i) = E_1(l_i) + \sum_{(k,i)\in A} M_{ki}^T$, and set relative beliefs $B_{ref}(l_i) = B(l_i) - B_{\min}(l_i)$, where $B_{\min}(l_i) = \min_{l_i \in N} B(l_i)$, B_{conf} is a threshold value, and let $Co(n_i) = \left|\left\{ l \in N : B_{ref}(l_i) \leq B_{conf} \right\}\right|$, then the priority of node n_i is:

$$\Pr(n_i) = \frac{1}{|Co(n_i)|} \qquad (6)$$

4.3 Label pruning and message scheduling

We discover there are more labels in the graph to be dealt with which result in the intolerable computational cost. While the label costs at all interior nodes are all equal to zero. This in turn implies that the beliefs at an interior node will initially be all equal as well, meaning that the node is "unconfident" about which labels to prefer. It suggests there is redundancy in the energy minimization procedure. Based on the thinking of reducing the number of labels from [2], a dynamic label pruning idea is proposed. If no label pruning, any message originating from a node will be very expensive to calculate. Afterward, a message scheduling based on priority is introduced. This scheme makes use of label pruning and allows us to always send cheap messages between the nodes of the graphical model. Through the two improvement methods, we can overcome the standard BP limitations of inefficiency to handle problems with a huge number of labels.

Each iteration procedure of priority BP is divided into a forward and a backward pass. The actual message scheduling mechanism as well as label pruning takes place during the forward pass. All nodes are visited in order of priority. We mark it as "committed" meaning that we must not visit him again during the current forward pass. The backward pass is then just to ensure that the other half of the messages gets transmitted as well. no label pruning takes place during this pass. A pseudo code description of Priority-BP introduced from [2] is shown as follows:

Algorithm 2: video completion with priority belief propagation

```
1  assign priorities to nodes and declare them
   uncommitted
   let $B_i^T = -(E_1(l_i) + \sum_{(k,i)\in A} M_{ki}^T)$
   for $k$ = 1 to $K$ do {$K$ is the number of iterations}
      execute ForwardPass and then BackwardPass
   assign to each node $i$ its label $\hat{l}_i$ that maximizes
   $B_i^k$
2  ForwardPass:
   for $t$ = 1 to $N$ do {$N$ is the number of nodes}
      $i$ = "uncommitted" node of highest priority
      apply "label pruning" to node $i$
      forwardOrder[$t$] = $i$; $i$→committed = true;
      for any "uncommitted" neighbor $j$ of node $i$ do
         send all messages $M_{ij}^t$ from node $i$ to node $j$
         update beliefs $B_j^t$ as well as priority of node $j$
3  BackwardPass:
   for $t$ = $N$ to 1 do
      $i$ = forwardOrder[$t$]; $i$→committed = false;
      for any "committed" neighbor $j$ of node $i$ do
         send all messages $M_{ij}^t$ from node $i$ to node $j$
         update beliefs $B_j^t$ as well as priority of node $j$
```

We dynamically reduce the number of possible labels for each node by discarding labels that are unlikely to be assigned to that node. Traditional BP algorithm is improved by our means, and experiment show the result is approving.

5 Experimental results

The algorithm described in this paper is applicable to a various full color natural videos of complex dynamic scenes such as DVD movie or home DV etc. With the assistance of our approach, common people can do many things that were only dealt with by artist or movie-maker early. As is well known, the perceived quality of the completed video frames

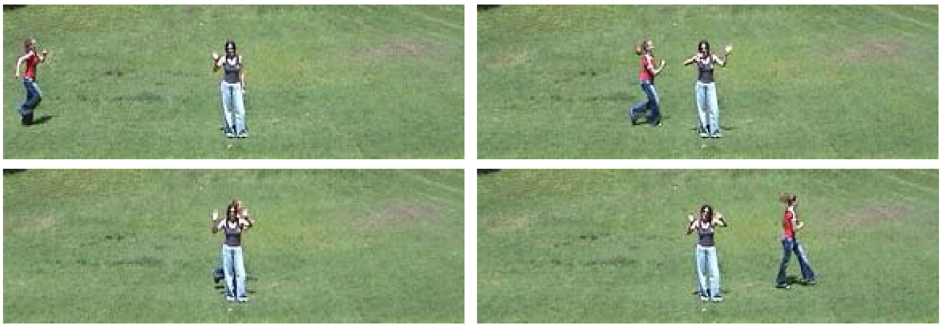


Fig. 1. The original image from video “hopping” containing total 240 frames. Top left to bottom right: frame 5, frame 86, frame 118, frame 163. Video data from [7].

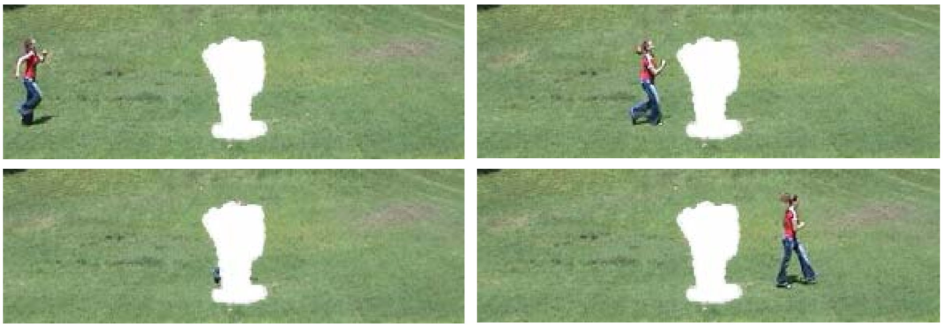


Fig. 2. Overview of the frames in which a person is removed from the original frames of Fig. 1. The blank region is the hole to be completed.

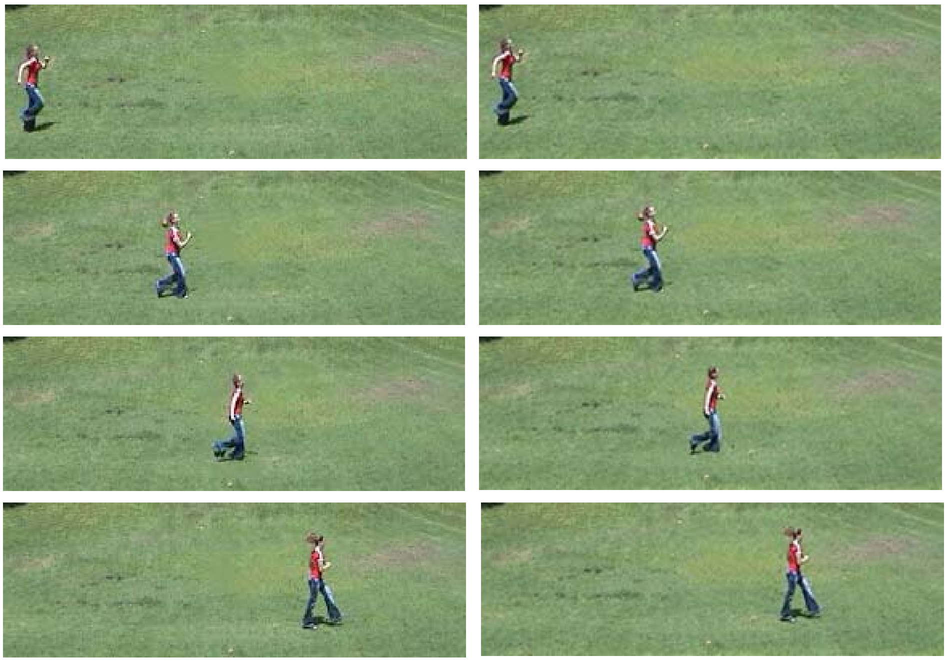
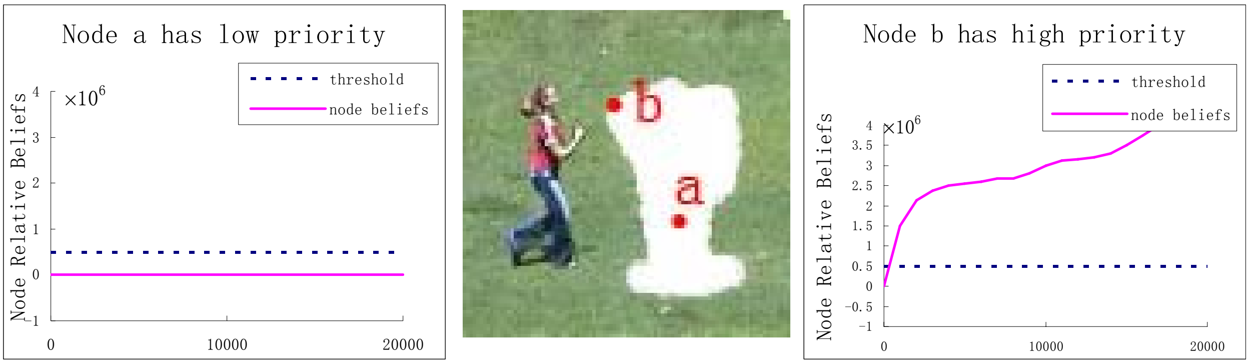


Fig. 3. Algorithm comparisons. Left column is the result from [7] and right column is the result from our approach.

depends on human perception, rather than mathematical measures, so we show some frames extracted from video sequences to demonstrate the effectiveness of our method. In this section, we mainly produce our algorithm result and the comparison with other ones. Video data used in this paper can be downloaded in: http://www.wisdom.weizmann.ac.il/~vision/VideoCompletion/.

Figure 1, 2 and 3 demonstrate the results from a two-woman video sequence after one woman has been removed. They show the completion performance with our method and Figure 3 indicates the comparison of the effect obtained by our belief



(a) node a has low priority (b) node a and b (c) node b has high priority

Fig. 4. node priority

propagation based method with the one obtained by other proposed methods. By means of the information from many presented frames, we can see clearly the portion of the moving object is absent due to the intersection with the unknown region. While using our method, the unknown portion can be completed very well even the hole is very large, and the result is very similarity to the original one.

Figure 4 shows the node priority. The relative beliefs of the interior node a are equal to zero, so it has lowest priority. While node b is visited early as its relative beliefs exceed the threshold.

In most cases, our method based on improved belief propagation algorithm could produce better effect comparison with previous methods.

6 Conclusions and future work

In this paper, a novel global optimization based method to fill the holes in a video sequence is proposed, which avoids visual inconsistent by greedy patch assignments. Through some definitions, a three-dimensional computational model is introduced for dealing with the video pixels. Aiming at the computational cost of standard belief propagation algorithm, a novel improved scheme named priority-BP has been proposed that carries two very important extensions over standard BP: label pruning and priority-based message scheduling. This method does not rely on specific prior knowledge and can be applied to any kind of video sequences. Experiment results indicate the result of our method is encouraging and present deep impressions for people.

If the camera is moving, our algorithm needs to be extended. We can make use of the motion parameters of the camera to obtain the knowledge of the background motion and estimate the projective homography between video frames. Our work using belief propagation can also be applied to more areas including 3d mesh completion. In the future, we will implement these extension works.

Acknowledgement

We would like to thank all that offer assistance to us. This research is supported by the National Grand Fundamental Research 973 Program of China (2002CB312106).

References:

[1] K. Bhat, M. Saptharishi, and P. Khosla. Motion Detection and Segmentation Using Image Mosaics. IEEE International Conference on Multimedia and Expo, vol. 3, pp. 1577-1580, July, 2000.

[2] N.Komodakis, G.Tziritas. Image completion using global optimization. In CVPR, 2006.

[3] Y. Li, J. Sun , H.-Y. Shum. Video object cut and paste, ACM Transactions on Graphics (TOG), v.24 n.3, July 2005

[4] P. Sand and S. Teller. Video Matching,. ACM Transactions on Graphics, vol. 22, no. 3, pp.592-599, July 2004.

[5] Y. Sugaya and K. Kanatani. Extracting moving objects from a moving camera video sequence. Proceedings of the 10th Symposium on Sensing via Imaging Information, pp. 279-284, June 2004.

[6] J. Sun, L. Yuan, J. Jia, and H.-Y. Shum. Image completion with structure propagation. In SIGGRAPH, 2005.

[7] Y. Wexler, E. Shechtman, M. Irani. Space-Time Video Completion. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04), vol 1, pp. 120-127, July 2004.

[8] Y. Zhang, J. Xiao, M. Shah. Motion Layer Based Object Removal in Videos. IEEE Workshop on Application on Computer Vision, Jan 5-6, Breckenridge, Colorado, 2005.