# Spatio-temporal grouping with constraint for seam carving in video summary application

Marc Décombas[AB]
[A]Laboratory MultiMedia Processing
Thales Communication & Security
Gennevilliers, France
marc.decombas@thalesgroup.com

Frédéric Dufaux[B], Béatrice Pesquet-Popescu[B]
[B]Signal and Image Processing Department
Telecom ParisTech/ Institut Télécom/LTCI
Paris, France
{dufaux,pesquet}@telecom-paristech.fr

*Abstract*—**Efficient tools to perform video summarization become essential as the number of cameras for video surveillance is growing exponentially. The development of automatic algorithms to aid human operators in identifying what is important and keeping it in a video summary becomes essential. This paper proposes a new way to obtain video summary based on seam carving. An efficient spatio-temporal grouping is done to determine the temporal rate of reduction depending on the content, to suppress groups of isolated seams, to identify spatio-temporal groups of seams and to approximate by constant segments the number of seams for each group, while keeping the total sum of seams constant. Problems of geometric distortion, anachronism and length of summary have been successfully addressed. The approach has been tested on video sequences of road traffic and leads to good visual results.**

***Keywords—Video summarization, video surveillance, seam carving, temporal saliency map***

## I. INTRODUCTION

In the last years, millions of video cameras have been deployed in the city’s streets, the highways and transportation hubs. In 2007, the numbers of surveillance cameras have reached 30 millions in the United States, producing over four billion of video footage each week [1]. This proliferation of cameras is due to an inexpensive camera network, easily deployed and remotely managed. Therefore, the development of automatic algorithms to aid human operators to identify what is important and store it in a video summary becomes essential.

This paper focuses on algorithms that compute a video summary for video surveillance applications. The video summary is an abbreviated video that preserves salient parts while removing spatio-temporal segments without interest. Such a summary is especially useful in video forensics, but can also be applied to quick video review for home applications [2]. Three main approaches to perform a video summarization have been proposed.

Firstly, with fast forwarding, frames are skipped in fixed or adaptive intervals. In the case of fixed intervals, salient objects that do not stay long enough are not visible. The approaches in [3][4][5] propose to skip frames at adaptive intervals. However, the limitation of these techniques is that, as only complete frames can be removed, the rate of temporal reduction, defined as the ratio of lengths of the original and processed videos, is relatively low.

A second approach to video summarization is to extract key frames and present them simultaneously as a storyline, as reviewed in [2]. Due to the fact that the keys frames can be very far away from each other, the temporal continuity of events is not preserved. Therefore, important contextual information is lost. An extension of this approach is to extract short sub-sequences, but the problem of having a relatively low reduction rate remains.

The two previous approaches preserve complete frames. The third approach is to extract salient spatio-temporal segments and combine them to obtain a video summary. In [6], Irani *et al.* combines spatial segments from different times to obtain a single image. A very high condensation of the content is obtained without losing any information but the dynamic aspects are lost. Based on the same idea, spatial segments can be shifted in time to obtain a summary. Kang *et al.* proposes in [7] to do video montage by modifying the location of spatial segments. This leads to a good rate of temporal reduction, but also to visible artifacts due to the combination of uncorrelated segments. In [8], a solution is proposed for panning camera where segments are aligned temporally and viewed simultaneously. Another approach of this work is proposed in [9][10][11], where an alternative to video montage called video synopsis is presented. Dynamic objects are identified, extracted and combined by using the minimization of a cost function. To prevent the total loss of context that can appear when spatio-temporal shifting is allowed, only temporal shift is allowed. The approach performs well, but the displacement of segments may cause a reversal of the order of activities. Another limitation of this approach is its complexity, as it involves several computationally complex tools such as object detection and background subtraction, clustering, and combination of spatio-temporal segments.

To cope with this puzzle, recent works propose to use an approach called seam carving. Seam carving has been introduced by Avidan in [12] to realize content aware image resizing. The idea is to suppress seams that are 8-connected paths defined vertically or horizontally from one side of the image to the other. These seams are computed on a saliency map that puts into evidence the important parts in the image. In [13], Rubinstein extends this approach to video applications and proposes a new way to compute the seams that take into account the influence of their suppression. A direct application of the seam carving without any constraint for producing the video summary creates temporal anachronisms, deformations of the objects and a summary not having the same length on all the lines. In [14][15], the authors propose to compute seams one by one with a spatio-temporal constraint. More specifically, the spatial constraint is the flexibility of the seam to move by one pixel to the left or to the right during its computation. The temporal constraint is the capacity of the spatial seam to evolve from one frame to the next one.

In this paper, we propose to perform video summary using seam carving with a spatio-temporal grouping constraint. Seam carving is applied in the (*x*,*t*) plane, with the aim of a reduction along the temporal dimension. Our approach first computes all the seams and then analyzes their evolution in space and time. We propose (1) a way to do an efficient spatio-temporal grouping that allow us (2) to determine a temporal rate of reduction in function of the content, (3) to suppress the group of isolated seams, (4) to identify sufficiently large spatio-temporal groups of seams, and (5) to approximate by constant segments the number of seams for each group, while keeping the total sum of seams constant. The proposed method avoids geometric deformations of the salient objects and anachronisms. In addition, the summary has the same length on all the lines. Our constraint enables more flexibility in the seam carving process, with seams that can better adapt to the content. At the same time, a better rate of temporal reduction can be achieved while preserving salient objects.

## II. TEMPORAL SEAM CARVING WITH GROUPS

Considering a video as a cube of dimensions (*N*,*M*,*T*), the seam carving can be applied in the (*x*,*t*) plane with a reduction along the temporal axis, thus leading to a summarized video.

Nevertheless, without any constraints, visual anachronisms and geometric deformations appear. Another problem is the number of seams suppressed on each line is not the same and the video has not the same length in function of the line. Previous approaches in [14][15] propose to add a spatio-temporal constraint directly during the seams computation. However, it does not allow a good adaptation of the seams to the content. Therefore, we propose to first compute all the seams in the (*x*,*t*) plane for all the *y*. Then, we apply our spatio-temporal grouping constraint. Before presenting the approach, a quick review of the seam carving is presented.

### *A.* Review of seam carving

Seam carving is an approach to resize images or video sequences while preserving the semantic content [12]. A seam is defined as an optimal – 8 connected path of pixels on a single image, going from top to bottom or from left to right. Formally, let *I* be an *N* x *M* image, a vertical seam is defined to be the set of points

$$s^X = s_i^x\ {}_{i=1}^{N} = x\ i\ , i\ {}_{i=1}^{N}, s.t. \forall i,\ x\ i\ - x\ i-1\ \leq 1$$

with *x* the horizontal coordinate of the point. A horizontal seam is defined similarly, by using the vertical coordinate. To define the seam, an energy function and a cumulative energy function are needed.

The energy function allows defining the salient parts of an image. Avidan and Shamir propose the gradient of the luminance [12]. The gradient is efficient to highlight the textured areas, although these areas are not necessarily salient. To address this problem, saliency map information has been integrated in the seam carving process, as in our previous work [16].

After defining the salient part in the energy function, it is necessary to define the optimal seam to remove. This is done by dynamic programming in the cumulative energy function. Avidan and Shamir first proposed the backward energy [12]. However, this function fails to measure the consequence of seam suppression and leads to some visual artifacts. Therefore, Rubinstein et al. proposed in [13] to use the forward energy, defined as:

$$M\ i,j\ = e\ i,j\ + \quad \begin{array}{l} M\ i-1, j-1\ + C_L(i,j) \\ M\ i-1, j\ + C_U(i,j) \\ M\ i-1, j+1\ + C_R\ i,j\ , \end{array}$$

where

$$\begin{aligned} C_L\ i,j\ &= I\ i,j+1\ - I(i,j-1) \\ &\quad + I\ i-1,j\ - I(i,j-1) \\ C_U\ i,j\ &= I\ i,j+1\ - I\ i,j-1 \\ C_R\ i,j\ &= I\ i,j+1\ - I(i,j-1) \\ &\quad + I\ i-1,j\ - I(i,j+1) \end{aligned}$$

where *e*(*i*,*j*) is an additional pixel based energy measure, for instance an energy function, and *I* the image.

### *B. Proposed approach*

Fig. 1 presents the proposed approach of video summarization using seam carving with spatio-temporal grouping constraint. From an original video of length *T*, saliency maps are created with the ST-RARE model [17] computed in the (*x*,*y*) plane. This model identifies the salient objects by finding the rarity on different maps. The most pertinent maps are combined together to obtain a unique saliency map. The original model uses static (L,a,b) and dynamic (amplitude and direction) components in order to identify salient areas for static and moving scenes. However, in the absence of movement, the model identifies static salient areas, preventing the suppression of the otherwise insignificant frames. Therefore, in our approach, only the dynamic component of the model is taken into account for video summarization.

To compute the seam, the forward energy of Rubinstein [13] is used on each frame. Contrary to [13] which uses a graph to define the seam and then force the seams to have a temporal link, dynamic programming is used on each frame separately in our approach. The seam carving with spatio-temporal grouping constraint is then applied on the (*x*,*t*) plane, giving a list of seams in each frame (*x*,*t*). These seams allowing to reduce *T* are suppressed in the original video to obtain a summarized version of length *T’* < *T*.

### C. *Rate of temporal reduction*

Seam carving is an iterative process allowing to pass from a resolution to another one depending on a stopping criteria. In most applications, seam carving is used to change the spatial resolution and the target output resolution is known a priori. In these cases, the stopping criterion is a vertical and horizontal number of seams to suppress.

In our case, the objective is to reduce as much as possible the temporal aspect while preserving the salient objects. For this purpose, a binarisation of the saliency map is used to separate the salient content that has to be preserved from the rest of the video. This binary saliency map is used as stopping criterion. While a seam does not cross the binary saliency map, the frame (*x*,*t*) passes to (*x*,*t*-1). Seam carving is applied independently on each frame (*x*,*t*) and stopped in function of the motion activity. As the objective is to preserve all the salient areas, and as some frames have more activity than others, the rate of temporal reduction is defined as the minimal number of suppressible seams. This step is the block "Uniformization of the number of seams" in Fig. 2.

### D. *Group of Seams*

If the seam carving is applied without any spatio-temporal constraint, some artifacts may appear as it can be observed in Fig. 6 (c), Fig. 7 (c), Fig. 8 (c) and Fig. 9 (c). The proposed approach solves this problem by creating groups of seams. In this way, the salient objects will shift by the same amount in time.

To perform the spatio-temporal grouping, the seams are first grouped together with a *Spatial_Distance* and a *Spatial_Threshold*. The *Spatial_Distance* is defined as

$$D\ Seam_t, Seam_{t+1}\ = \max_{i=1\ldots N}\ Seam_t\ i\ , Seam_{t+1}\ i$$

where *N* is the length of a seam. This distance has been chosen as it successfully identifies salient objects between seams, contrary to the mean or the median. The Spatial_Threshold represents the maximal distance between two consecutives seams found in the same group. It has been experimentally defined at 7 pixels. Then, the groups have to be temporally linked together. For this purpose, the symmetric difference is used between the groups of seams areas at *t* and at *t*+1. The symmetric difference is defined as the area of the union minus the area of the intersection. The area of the groups of seams is bordered by the most left seams and the most right seams. Then, the temporal regrouping is done in function of the minimum of this difference. These steps are included in the block "spatio-temporal grouping" of Fig. 2.

Next, the groups of seams are processed as illustrated in Fig. 3 to solve the problem of geometric distortions and anachronism. The number of seams by group varies with the time ($\mathrm{Nb\_Seam}_{\mathrm{Gpe\_x}}=\mathrm{f}_{\mathrm{Gpe\_x}}(\mathrm{t})$). Isolated groups, not present in enough frames, are suppressed. The associated number of seams is reallocated to other groups. Then, a median filter is applied temporally on each function $\mathrm{f}_{\mathrm{Gpe\_x}}$ to suppress strong local variations. To avoid anachronism, the function $\mathrm{f}_{\mathrm{Gpe\_x}}$ has to be piece-wise constant. The length of the pieces is linked to the size of the salient objects and the time they remain in the scene. Rupture detection is done to segment the function in pieces. The rupture detection is based on the variation of the function compared to the median. The median value is associated to each piece. A set of constant pieces is then obtained for each group.

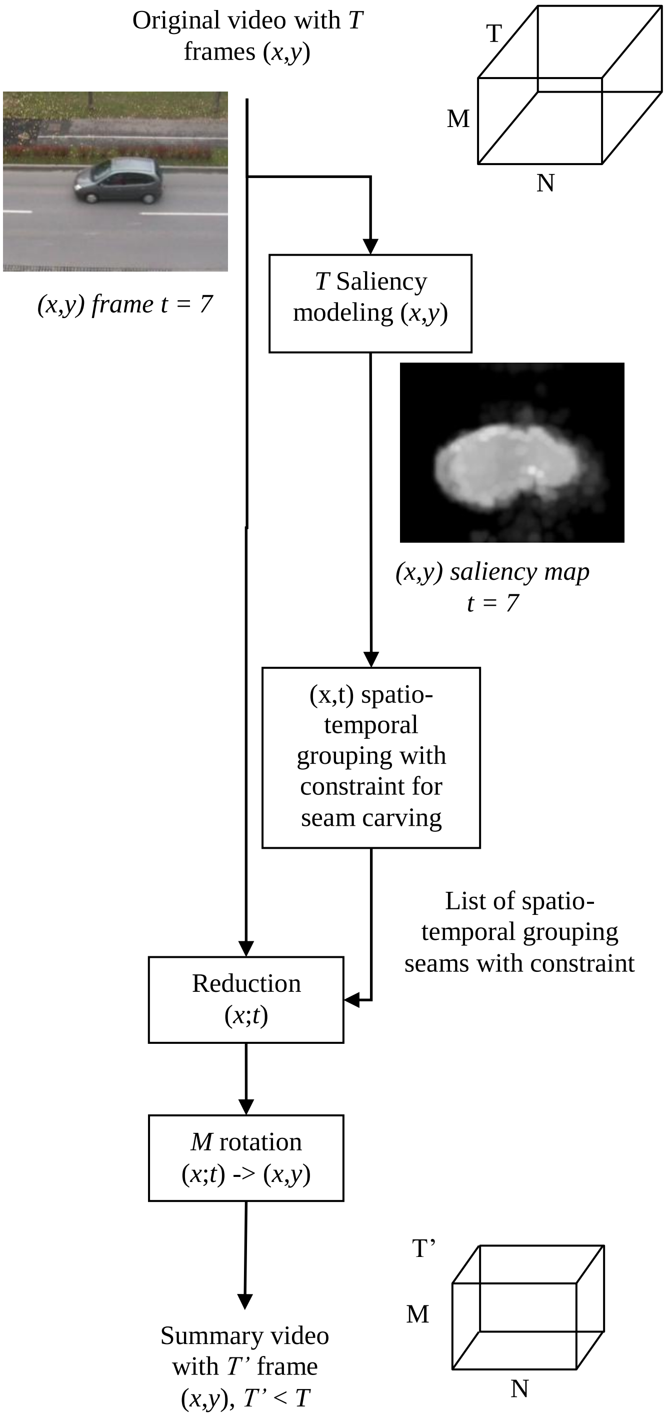

Original video with *T* frames (*x*,*y*)
T
M
N
*(x,y) frame t = 7*
*T* Saliency modeling (*x*,*y*)
*(x,y) saliency map t = 7*
(x,t) spatio-temporal grouping with constraint for seam carving
List of spatio-temporal grouping seams with constraint
Reduction (*x*;*t*)
*M* rotation (*x*;*t*) -> (*x*,*y*)
Summary video with *T'* frame (*x*,*y*), *T' < T*
T'
M
N

Fig. 1 Proposed approach of video summary with spatio-temporal grouping constraint for seam carving.

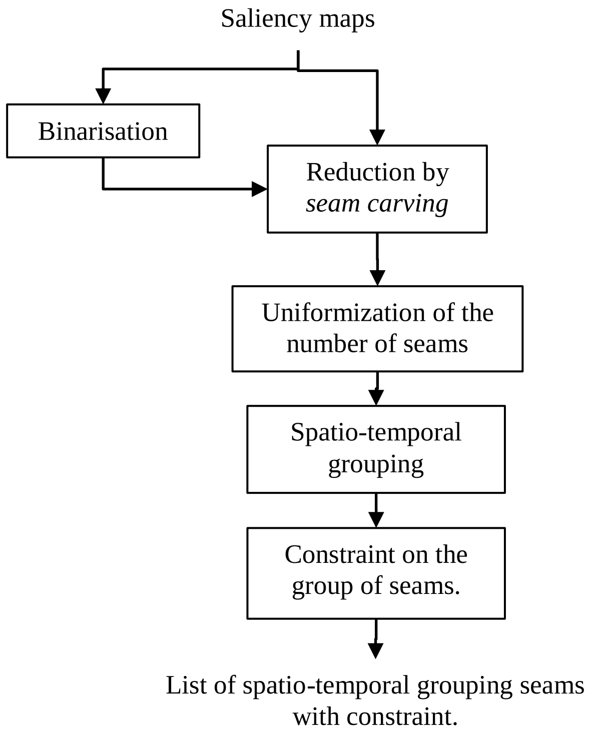

Saliency maps
Binarisation
Reduction by *seam carving*
Uniformization of the number of seams
Spatio-temporal grouping
Constraint on the group of seams.
List of spatio-temporal grouping seams with constraint.

Fig. 2 Spatio-temporal grouping with constraint for seam carving.

## III. EXPERIMENTS AND VISUAL RESULTS

To evaluate the proposed method, video surveillance test sequences have been chosen with a fixed camera and salient objects (cars, bicycles) crossing the scene from left to right or from right to left. In addition, many frames do not contain any activity. This type of sequences is very representative of video surveillance applications.

In Fig. 4, the suppressible seams (in red) can be seen after the permutation for $y = \{139,154, 244, 253\}$. The road is in grey and the vehicle trajectories are in black or white depending on the vehicles colors. On the images (a) and (b), the trajectories of 4 vehicles going from left to right on the top of the road are visible. On the images (c) and (d), the trajectory of one vehicle going from right to left on the bottom of the road is visible. With our approach, the quantity of seams suppressed between the vehicles is constant, as it can be seen between $y$ = 139 and $y$ = 154 or $y$ = 244 and $y$ = 253. The seams can adapt to the number of salient objects and their trajectories, as it can be seen on all the frames $(x,t)$ of Fig. 4.

The suppressible seams from the Fig. 4 on the $(x,t)$ plane for $y = \{139,154,244,253\}$ can also be seen in Fig. 5 for $t = \{17,40,58\}$. The seams avoid the salient objects at $t=\{17,58\}$ and when there is no activity in the frame, the frame is totally suppressed, as at $t$ = 40.

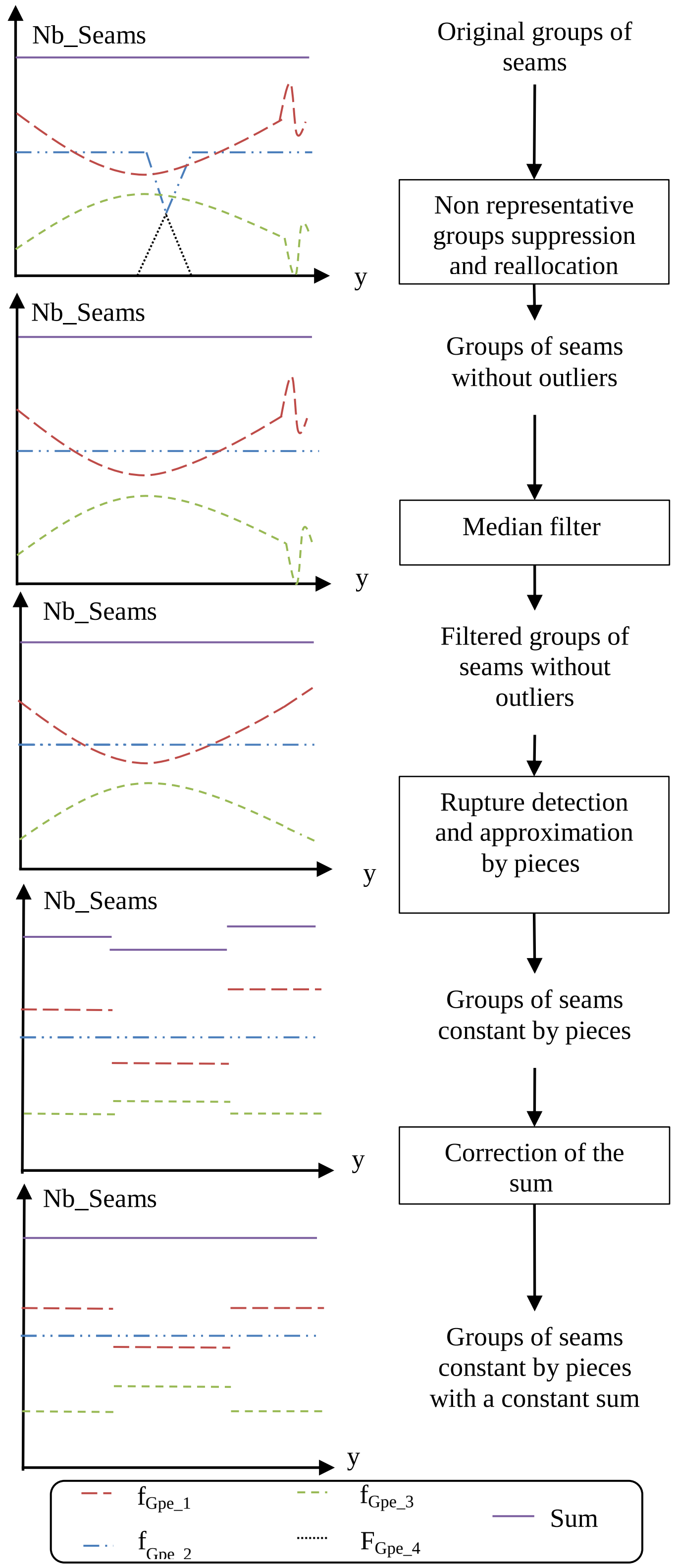

Nb_Seams
y
Nb_Seams
y
Nb_Seams
y
Nb_Seams
y
Nb_Seams
y
Original groups of seams
Non representative groups suppression and reallocation
Groups of seams without outliers
Median filter
Filtered groups of seams without outliers
Rupture detection and approximation by pieces
Groups of seams constant by pieces
Correction of the sum
Groups of seams constant by pieces with a constant sum
$f_{Gpe\_1}$
$f_{Gpe\_2}$
$f_{Gpe\_3}$
$F_{Gpe\_4}$
Sum

Fig. 3 Processing applied on group of seams. From original groups to piece-wise constant groups.

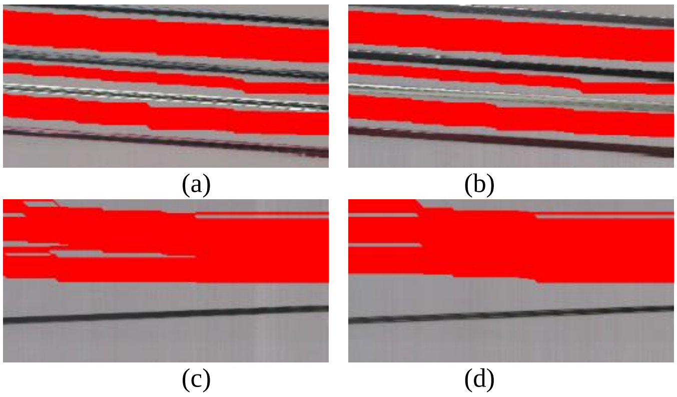


Fig. 4 Visualization of the *(x,t)* plane for video-A (a) y=139, (b) y=154, (c) y = 244, (d) y=253; the suppressible seams are in red, the road is in grey, and the different vehicle trajectories are in black or white.

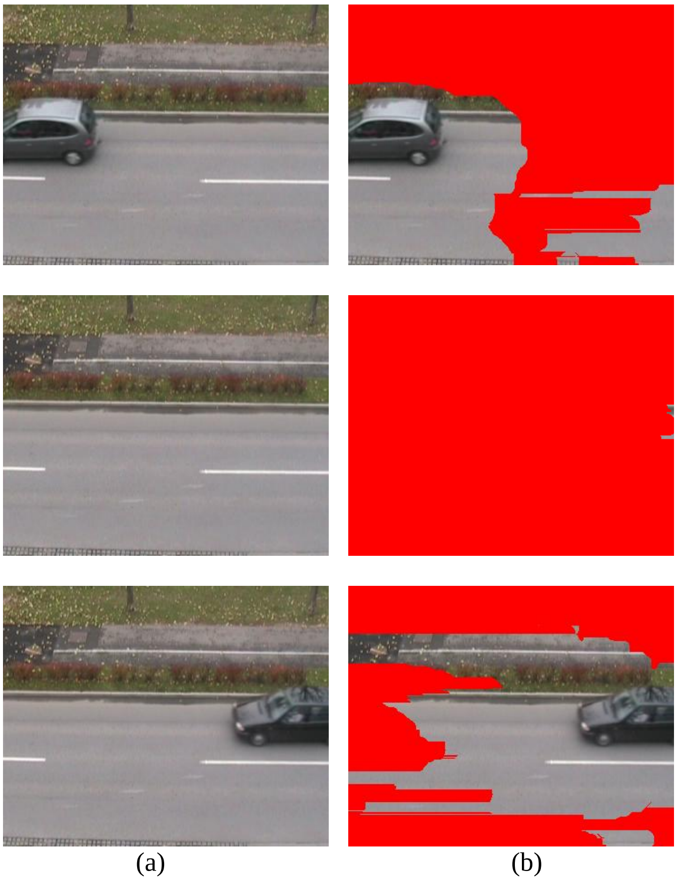
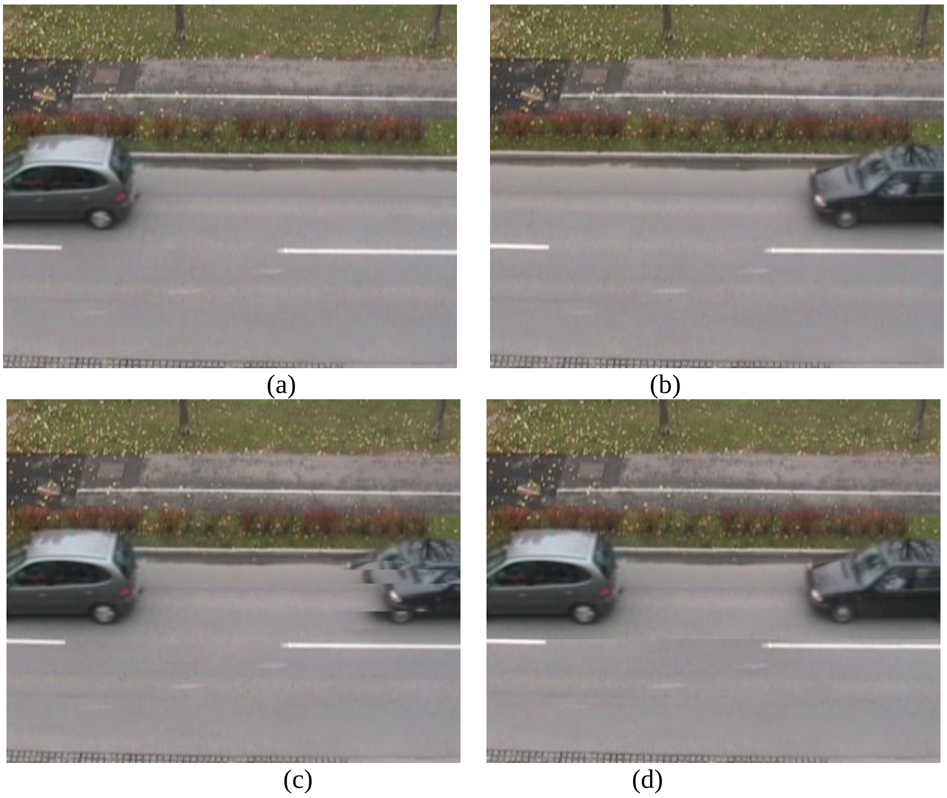


Fig. 5 Visualization of the original frame for video-A (a) at *t* = 17, 40, 58 and the frame with the suppressible seams (b) at *t* = 17,40,58

Results of our approach are shown in Fig. 6, Fig. 7, Fig. 8, and Fig. 9. In Fig. 6 (a) and Fig. 6 (b), vehicles from the original video at *t* = 17 and *t* = 58 are shown. On Fig. 6 (c) the approach to do temporal summary based on seam carving without constraint is illustrated. The second vehicle arrives earlier but some geometrical artifacts have appeared. This is due to the fact that the number of seams suppressed before and after the vehicle is not constant on all the lines. As a consequence, pieces of the vehicle arrive earlier than the rest of the vehicle. In our approach, all the lines of the second vehicle are shifted by the same quantity. The vehicle has no artifacts, as can be seen in Fig. 6 (d). In Fig. 7 (a), (b), two vehicles are visible at *t* = 77 and *t* = 94 in the original video. In Fig. 7 (c), the approach without constraint deforms the shape of the vehicles, which now appear shifted in time at *t* = 38. With our approach in Fig. 7 (d), the same temporal shifting is obtained but the vehicles are well preserved.

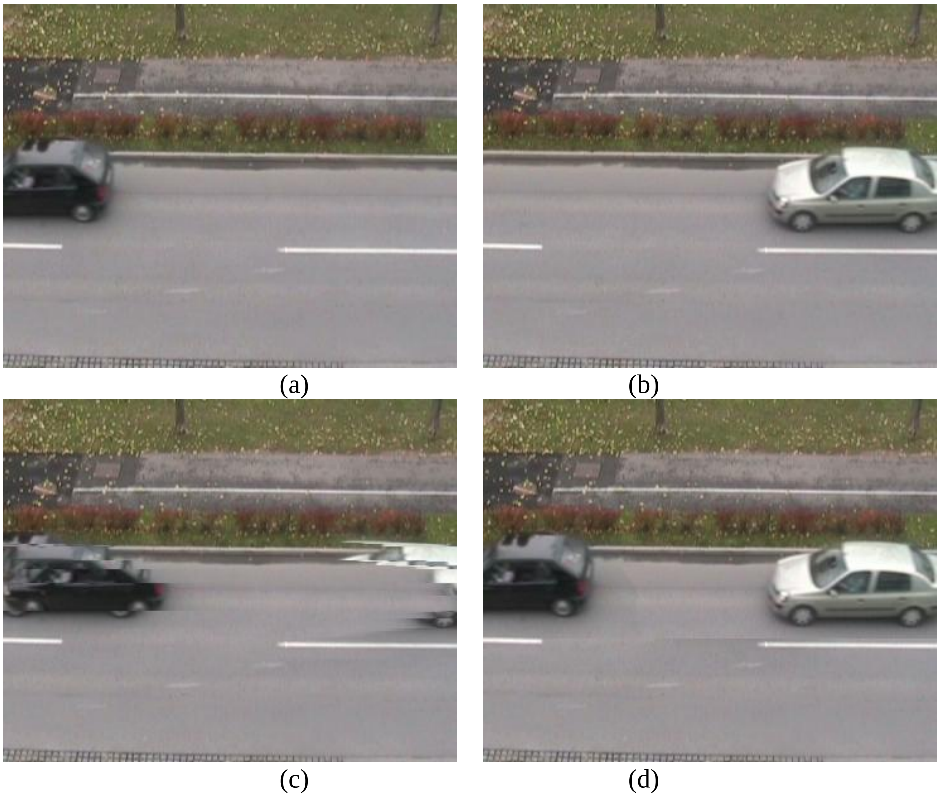
Similar results are shown in Fig. 8 and Fig. 9. Artifacts are visible on the road in Fig. 8 due to a temporally changing luminance. In all cases, the seam carving without constraint is clearly less efficient than the proposed approach. The video-A has 35 frames without activity and the process allows to suppress 68 frames on the total sequence of 180 frames. The video-B has 60 frames without activity and the process allows to suppress 77 frames on the total sequence of 140 frames. Finally, the video-C has 52 frames without activity and the process allows us to suppress 66 frames on the total sequence of 120 frames.



(a) (b)

(c) (d)

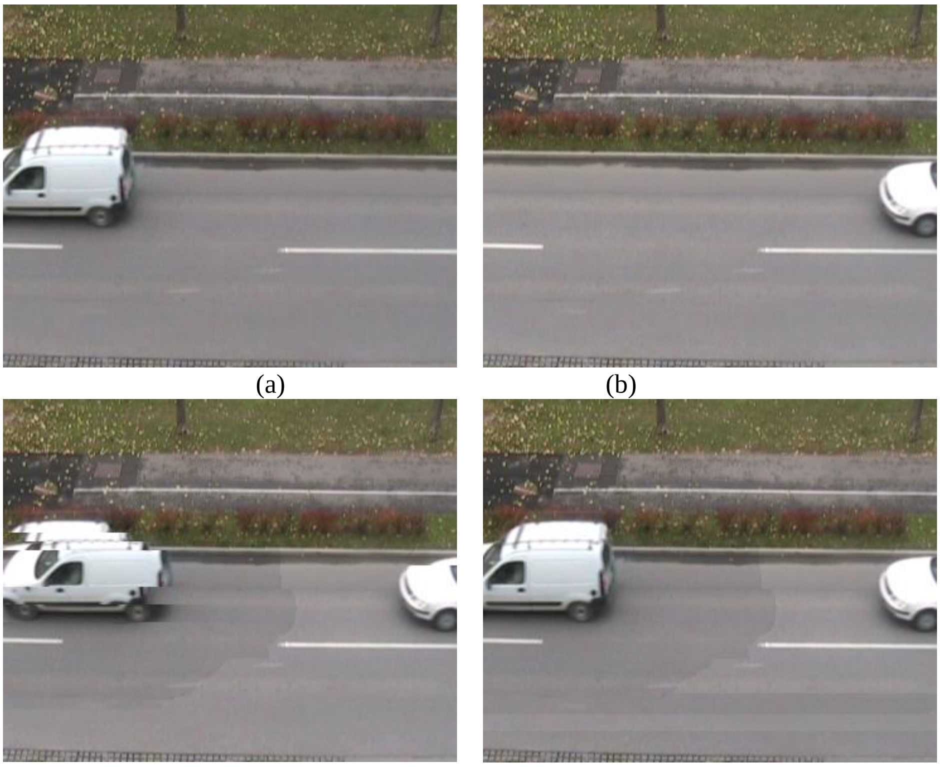
Fig. 6 Video-A summary: (a) original frame at t = 17, (b) original frame at t = 58, (c) frame after seam carving without constraint at t= 17, (d) frame after spatio-temporal grouping with constraint for seam carving at t=17



(a) (b)

(c) (d)

Fig. 7 Video-A summary: (a) original frame at t = 77, (b) original frame at t = 94, (c) frame after seam carving without constraint at t= 38, (d) frame after spatio-temporal grouping with constraint for seam carving at t=38



(a) (b)

(c) (d)

Fig. 8 Video-B summary: (a) original frame at t = 34, (b) original frame at t = 106, (c) frame after seam carving without constraint at t= 29, (d) frame after spatio-temporal grouping with constraint for seam carving at t=29

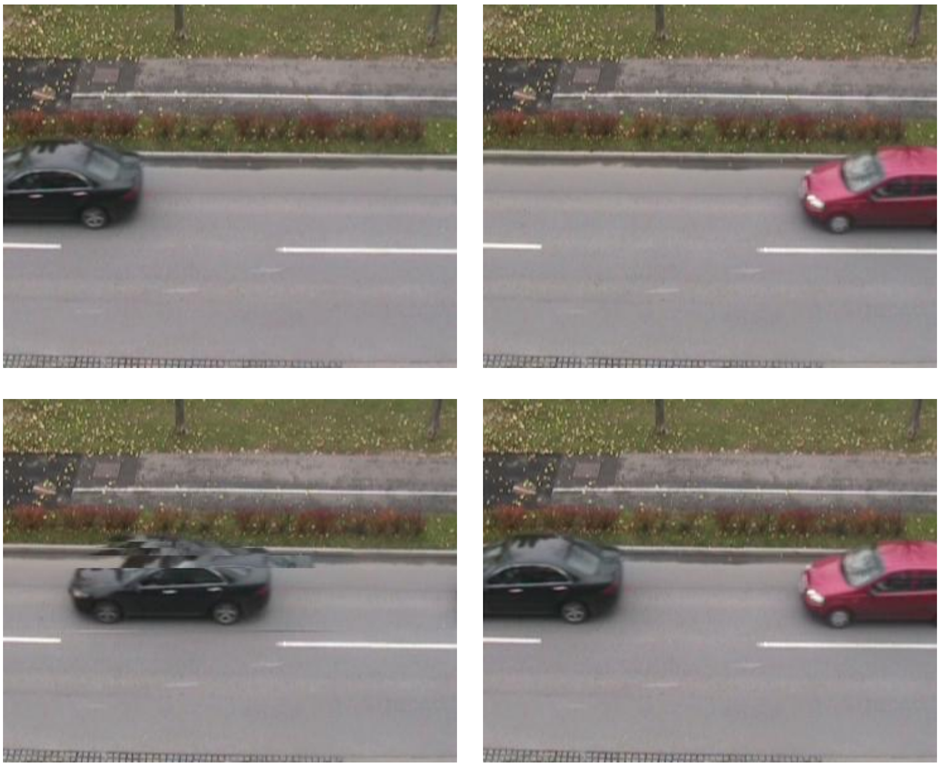


Fig. 9 Video-C summary: (a) original frame at t = 29, (b) original frame at t = 85, (c) frame after seam carving without constraint at t= 20, (d) frame after spatio-temporal grouping with constraint for seam carving at t=20

## IV. CONCLUSION AND FUTURE WORK

In this paper, a seam carving with spatio-temporal grouping constraint for video summary applications has been presented. Our proposed approach achieves a temporal rate of reduction adapted to the content, suppressing and reallocating the groups of isolated seams. An identification of important spatio-temporal groups of seams and an approximation of the number of seams by piece-wise constant segments has been introduced. In this way, the problems of salient objects geometric deformation and anachronisms have been solved. Without any constraint, the number of seams suppressed on each line is not necessary the same. With our approach, this number is defined in function of the content and the summary has the same length for all the lines. Our approach yields good results due to the fact that the constraints on the number and spatio-temporal position of suppressible are applied after having obtained the seams. This lets more flexibility to the seams and leads to a good rate of reduction, while preserving the content.

Future work will be to automatically define the spatial and the temporal thresholds on the distances, have a multi dimension detection of rupture and a correlated approximation of the number of seams by groups.

## REFERENCES

[1] J. Vlahos, "Welcome to the planopticon," *Popular Mechanics.*, vol. 1, no. 1, pp. 64–69, Jan.2008 .

[2] J. Oh, Q. Wen, J. Lee, and S. Hwang, "Video abstraction," in *VideoData Mangement and Information Retrieval*, S. Deb, Ed. Hershey, PA: Idea Group, Inc./IRM Press, pp. 321–346, chap. 3, 2004.

[3] M. Yeung, and B.-L.Yeo, "Video visualization for compact presentation and fast browsing of pictorial content," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 5, pp. 771–785, Oct. 1997.

[4] J. Nam, and A. Tewfik, "Video abstract of video," in *IEEE Proc. On MultiMedia Signal Processing*, pp. 117–122, Copenhagen, Denmark, 1999.

[5] N. Petrovic, N. Jojic, and T. Huang, "Adaptive video fast forward," *Multimedia Tools Appl.*, vol. 26, no. 3, pp. 327–344, Aug. 2005.

[6] M. Irani, P. Anandan, J. Bergen, R.Kumar, and S. Hsu, "Efficient representations of video sequences and their applications," *Image Commununication Signal Procesing*, vol. 8, no. 4, pp. 327–351, May 1996.

[7] H.-W. Kang, Y. Matsuhita, X. Tang, and X.-Q. Chen, "Space-time video montage," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, , vol.2, pp. 1331–1338, Jun. 2006.

[8] A. Rav-Acha, Y. Pritch, D. Lischinski, and S. Peleg, "Dynamosaicing: Mosaicing of dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1789–1801, Oct. 2007.

[9] A. Rav-Acha, Y. Pritch, and S. Peleg, "Making a long video short: Dynamic video synopsis," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*,vol.1, pp. 435–441, Jun. 2006.

[10] Y. Pritch, A. Rav-Acha, and S. Peleg, "Non-chronological video synopsis and indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, Nov. 2008.

[11] Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg, "Clustered Synopsis of Surveillance Video", *6th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Genoa, Italy, Sept. 2009.

[12] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graph.*, vol. 26, no. 3, pp.10, 2007.

[13] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargetting," *ACM Trans. Graph.*, vol. 27, no. 3, pp.16, 2008.

[14] B. Chen and P. Sen, "Video carving," presented at the UROGRAPHICS, Crete, Grece, 2008.

[15] Z. Li, P. Ishwar and J. Konrad, "Video Condensation by Ribbon Carving", *IEEE Trans. on Image Processing*, vol.18, pp. 2572 – 2583, 2009.

[16] M. Décombas, F. Dufaux, E. Renan, B. Pesquet-Popescu, F. Capman, "Improved seam carving for semantic video coding",*IEEE Proc. On MultiMedia Signal Processing*, Banff, AB, Canada, Sept. 2012.

[17] M. Décombas, N. Riche, F. Dufaux, B. Pesquet-Popescu, M. Mancas, B. Gosselin, and T. Dutoit, "Spatio-temporal saliency based on rare model", *IEEE Proc. Int. Conf. on Image Processing*, Melbourne, Australia, Sept. 2013 – submitted.