Space-Time Video Montage

Hong-Wen Kang* Yasuyuki Matsushita† Xiaoou Tang† Xue-Quan Chen

University of Science and Technology of China
Hefei, P.R.China
{hwkang@mail.,chenxq@}ustc.edu.cn

Microsoft Research Asia†
Beijing , P.R.China
{yasumat,xitang}@microsoft.com

Abstract

Conventional video summarization methods focus predominantly on summarizing videos along the time axis, such as building a movie trailer. The resulting video trailer tends to retain much empty space in the background of the video frames while discarding much informative video content due to size limit. In this paper, we propose a novel space-time video summarization method which we call space-time video montage. The method simultaneously analyzes both the spatial and temporal information distribution in a video sequence, and extracts the visually informative space-time *portions of the input videos. The informative video portions are represented in volumetric layers. The layers are then packed together in a small output video volume such that the total amount of visual information in the video volume is maximized. To achieve the packing process, we develop a new algorithm based upon the first-fit and Graph cut optimization techniques. Since our method is able to cut off spatially and temporally less informative portions, it is able to generate much more compact yet highly informative output videos. The effectiveness of our method is validated by extensive experiments over a wide variety of videos.*

1. Introduction

The rapid increase of the amount of on-line and off-line video data necessitates development of efficient tools for fast video browsing. Video summarization [6, 14, 13] is one approach toward tackling this problem, in that it automatically creates a short version of the original input video. Summarized video content is important for many practical applications such as archiving 24-hour security videos and providing easy access to large sets of digital video documentaries.

This paper addresses the problem of automatically synthesizing a new short/small video from a long input video sequence by extracting and fusing the *space-time* informative portions of the input video. Unlike prior video summarization methods, our method is not limited to the frame-basis, but uses an arbitrary space-time volume that is extracted as the informative video portion. The extracted 3D informative video portions are packed together in the output video volume in a way in which the total visual information is maximized. This approach generates a compact yet highly informative video that tries to retain most informative portions of the input videos.

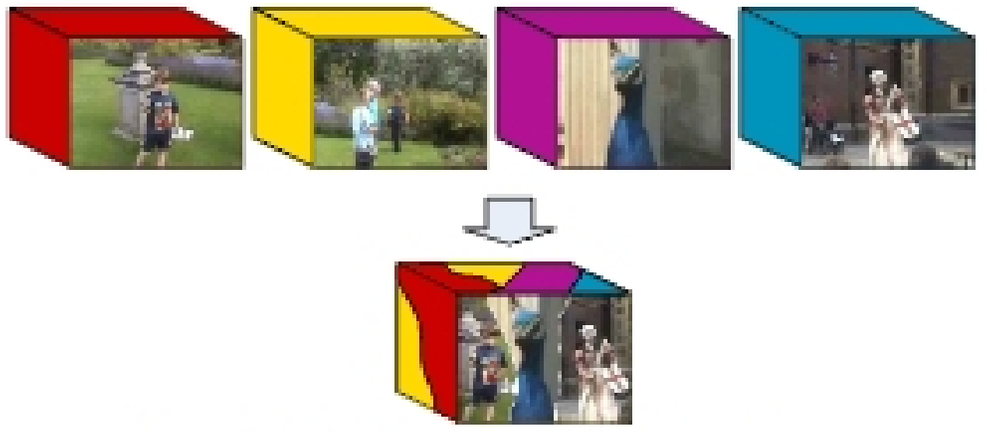

Figure 1. Idea of the space-time video montage.

Prior works on video summarization [6, 14, 13] have typically been based on a two-step approach. First, video streams are divided into a set of meaningful and manageable segments called shots. Then key frames were selected according to criteria from each video shot to generate a summary video. Although these approaches can extract some basic information of the video, they suffer a common disadvantage. They are all frame-based, i.e. they treat a video frame as a non-decomposable unit. Therefore, the resulting video tends to appear to be a fast-forward version of the original video while retaining a large amount of empty space in the video frame background.

Our approach is built upon the idea that some space-time video portions are more informative than others. Considering that visual redundancy exists in videos, the assumption is apparently correct. However the definition of "informa-

*This work was done while the first author was visiting Microsoft Research Asia.

0-7695-2646-2/06 $20.00 (c) 2006 IEEE

tive" is not straightforward since it involves the problem of image understanding. There has been extensive work aimed at extracting salient image/video parts [11, 8, 15]. In general, video summarization methods try to determine important video portions while relying on studies of pattern analysis and recognition.

In our method, we extract and represent space-time informative video portions in volumetric layers. The idea of layered representations was introduced by Wang *et al.* [16] in computer vision and has been widely used in many different contexts [2, 17]. The layered representation has often been used for describing regional information such as foreground and background or different objects with independent motion. Here, we use the layered representation for depicting saliency distribution. A layer is assigned to each high-saliency video portion, each of which represents a different saliency distribution.

1.1. Proposed approach

We try to develop an effective video summarization technique that can generate a highly informative video, in which the space-time informative portions of the input videos are densely packed.
This paper has three major contributions:
Space-time video summarization. Our method treats the informative video portions as a space-time volume without being limited by a per-frame basis. It allows us to develop a new video summarization method which can generate compact yet highly informative summary videos.
Layered representation of informative video portions. We propose an idea of representing informative video portions in the form of volumetric layers such that each layer contains an informative space-time video portion. We call the volumetric layer a saliency layer. The saliency layers are used to compute the optimal arrangement of the input video portions for maximizing the total amount of information in the output video.
Volume packing and merging algorithm. To achieve the goal of packing the informative video portions, we develop a new 3D volume packing and merging algorithm based upon the first-fit algorithm [5] and Graph cut [4] optimization technique. Using this method, the saliency layers are packed and merged together in the output video volume to generate the final montaged video.

In the rest of the paper, we first formulate the problem of space-time video montage in Sec. 2. In Sec. 3, we describe the detailed algorithm of our space-time video montage method. We show the experimental results in Sec. 4 followed by conclusions in Sec. 5.

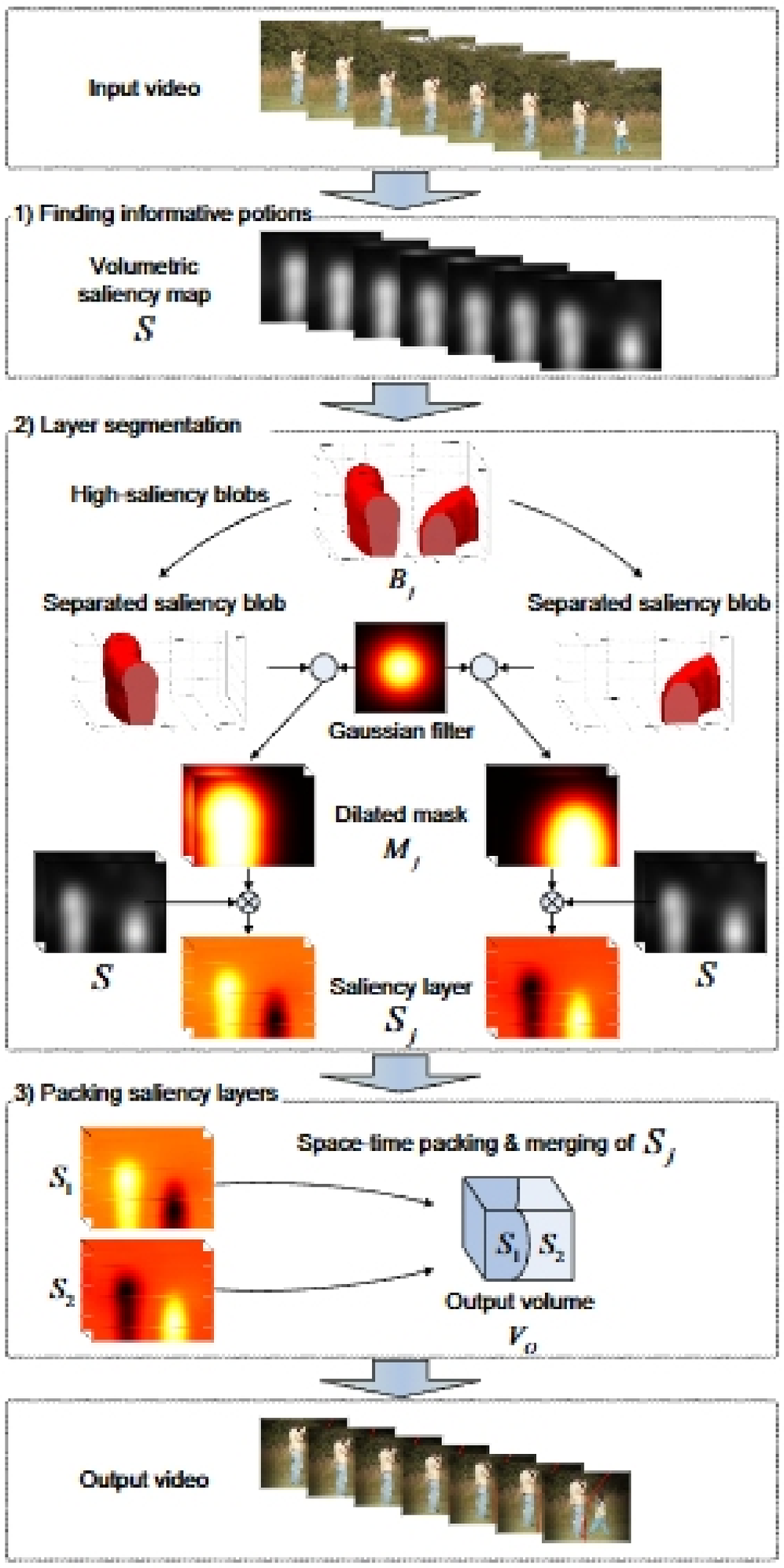

Figure 2. Overview of the space-time video montage.

2. Overview of Space-time Video Montage

The problem of space-time video montage consists of three sub-problems, i.e., finding informative video portions, layer segmentation of informative video portions and packing them in an output video volume. In this section, we present an overview of the problem of space-time video montage and notations which are used in the rest of the paper.
Finding informative video portions. The first problem in space-time video summarization is finding infor-

mative video portions from the long input video sequence V. Defining the amount of information is a difficult problem since it requires image understanding. There exist many methods that try to extract saliency from images/videos [11, 8, 15]. The actual implementation of our saliency measure will be detailed in Sec. 3. Supposing that we are able to assign saliency values to all the video pixels, we are able to obtain a *saliency volume* S that is associated with the input video volume V.

Layer segmentation. The saliency volume S may contain a number of isolated informative portions where high-saliency values are assigned. Here we introduce the idea of *saliency layers* to separately treat those informative portions. The saliency layers $\mathcal{S} = \{S_j : j = 1, \ldots, n\}$ are extracted from the original saliency volume S, where n is the number of layers. We use the notation S_j to represent the j-th layer.

Packing saliency layers. The problem of packing salient video portions into an output video volume such that the total saliency value grows to its maximum and can be viewed as a variant of the *Knapsack problem* [7], which is a classic combinatorial optimization problem. The goal of the Knapsack problem is to pack a set of items into a limited size container such that the total importance of items becomes maximum. Although our problem is similar to the Knapsack problem, the following differences exist: input items are video volumes, each of which can have a larger volume than the output volume; every video pixel in the video volumes is associated with its importance; and input items can overlap each other.

Denoting the output video volume as V_o and the associated saliency volume as S_o, our goal is to pack the input video volume V into the output video volume V_o in a way that S_o contains maximal saliency from $\mathcal{S}$. It is equivalent to finding the optimal space-time translations $\mathbf{x}_j$ of the saliency layers S_j which maximizes the following objective function:

$$\sum_{\mathbf{p} \in S_o} f(S_j(\mathbf{p} - \mathbf{x}_j)), \tag{1}$$

where $f(\cdot)$ is the function which evaluates the saliency value for each pixel $\mathbf{p} = (x, y, t)^T$. For instance, $f(\cdot)$ can be defined as $f(\cdot) = \max_j(\cdot)$ which takes the maximum saliency value at a pixel where the saliency layers are overlapped. Since the saliency layers are bounded by the original input video volume, it follows $S_j(\mathbf{x}) = 0$ if $\mathbf{x} \notin S_j$. Once the positions $\mathbf{x}_j$ are determined, the color values of the output video V_o are assigned by composing the input video according to the arrangement of the saliency layers. In the case of $f(\cdot) = \max_j(\cdot)$, for instance, by denoting $V(\mathbf{p})$ to represent the color value at the pixel $\mathbf{p}$ in the video volume V, a simple form of the composition can be described as

$$V_o(\mathbf{p}) = \{V(\mathbf{p} - \mathbf{x}_j) : j = \arg\max_j(S_j(\mathbf{p} - \mathbf{x}_j))\}. \tag{2}$$

In the following sections, we describe the implementation details to solve this problem.

3. Implementation

In this section, we describe the details of the algorithm. The overview of the proposed method is illustrated in Fig. 2. Our algorithm consists of three major stages: (1) finding informative video portions, (2) layer segmentation of the saliency volumes and (3) packing saliency layers. In the following subsections, we describe the implementation details of each stage.

3.1. Finding informative video portions

In order to determine salient portions in video, we define a spatio-temporal saliency measure using the spatio-temporal contrast. Our spatio-temporal saliency $S(\cdot)$ at a video pixel position $\mathbf{p}$ is defined using the neighboring pixels $\mathbf{q} \in \mathcal{Q}$ as

$$S(\mathbf{p}) = \mathcal{GS}\{\sum_{\mathbf{q} \in \mathcal{Q}} d_S(\mathbf{p}, \mathbf{q})\}, \tag{3}$$

where the distance function d_S denotes the stimulus measure and $\mathcal{GS}$ is a Gaussian smooth operator with $\sigma = 3.0$. We define d_S as the l^2-norm color distance:

$$d_S(\mathbf{p}, \mathbf{q}) = \|\mathbf{I}(\mathbf{p}) - \mathbf{I}(\mathbf{q})\|_2, \tag{4}$$

where $\mathbf{I}(\cdot)$ is the color vector in the LUV space.

Once the saliency values are computed for all of the pixels in the video volume, the saliency values are normalized to the range of $[0, 1]$.

3.2. Layer segmentation of saliency volumes

In the original volumetric saliency map S, there exist a set of high saliency portions in a low-saliency background. In order to treat the high saliency portions separately, we perform layer segmentation of the volumetric saliency map so that each layer only contains one salient portion. The layer segmentation consists of three steps: (a) segmentation of high saliency portions, (b) morphological growing of saliency portions and (c) assignment of negative saliency values.

(a) Segmentation of high saliency portions. The first stage of the layer segmentation is generating *saliency blobs* that represent the high-saliency portions in the input video. To locate the high-saliency portions and separate them from