REALTIME DETECTION OF SALIENT MOVING OBJECT: A MULTI-CORE SOLUTION

Patricia P. Wang, Wei Zhang, Jianguo Li, Yimin Zhang

Intel China Research Center
8F, Raycom Infotech Park A, Beijing, P.R.China, 100080

ABSTRACT

Detection of salient moving object has great potentials in activity recognition, scene understanding, etc. However techniques to characterizing the object in fine granularity have not been well developed in real applications due to the computational intensity. The emerging multi-core technology in hardware design provides an opportunity for the compute intensive algorithms to boost speed in parallel. This paper proposed a scalable approach to detecting salient moving object which is designed inherently for parallelization. To characterize the object in fine granularity, we extract color-texture homogenous regions as the basic processing unit by image segmentation. To identify salient object, we generate probabilistic template by learning the space-time context. The parallel algorithm is implemented using OpenMP. Evaluations have been carried out on sports, news, and home video data. For the CIF size image, we get processing speed of 51.1 frames per second and near linear speed up on an eight-core machine. It indicates that the algorithm parallelization is a promising solution for practical applications in the multimedia field.

Index Terms— Salient object, space-time context, unsupervised learning, parallel processing

1. INTRODUCTION

In the field of video mining, the task of high-level feature extraction has attracted tremendous efforts for the past decades. Among them, identifying salient object has large potentials in the applications of activity recognition, scene understanding, etc. Most attempts focus on still images [1, 2, 3, 4], while some works on motion pictures [5]. Roughly, salient object detection can be categorized into three classes: bottom-up, top-down, and hybrid combination. For the bottom-up methods [1, 2, 4], the representation granularity of object depends on what basic processing unit, e.g. pixel, grid, boundary, is extracted. For the top-down methods [5], great challenges exist in the space-time modeling due to the diverse appearance of object, e.g. shape, color, texture, motion. For the hybrid combination methods [3], the fusion strategy, e.g. feature-, model-, decision-level, is crucial to balance the detection accuracy and algorithm generalization.

For the advanced applications in video mining, such as activity recognition, it is highly demanding to characterize the salient moving objects in their natural shapes. The development of such techniques however has been limited due to two main issues: i) it is too computationally expensive to utilize them in real applications, if we extract the object boundary and shape for every video frame; ii) it is quite challenging to distinguish salient moving object from other uninteresting regions in a totally automated way. In this situation, the computation intensity is the primary issue preventing the fine-grained detection approach being developed from real applications. We have noticed that a great change in the architecture of hardware is on the way from dual-, quad-, eight-core to tens or even hundreds of cores. If shift from serial-based thinking to parallelism [6] in the algorithm design and implementation, it is possible to boost speed by taking advantage of the multi-core architectures.

This paper proposed an approach to detecting and characterizing the salient moving object in fine-grained shape and boundary. It is a bottom-up algorithm with unsupervised modeling of space-time context. The approach consists of three main phases: first, segmenting the video frame into color-texture homogenous regions, and extracting motion features from adjacent frames; second, cleaning noises in the image segmentation results, and merging raw regions into interesting and uninteresting region clusters in terms of motion intensity and direction; and third, modeling the space-time context with a probability template, and verifying the salient object from a statistics perspective. To solve the two main issues that preventing the fine-grained detection approach from real applications, i) we designed an easy to parallel algorithm, so that it is scalable with the increasing number of threads (cores); ii) we presented an unsupervised learning algorithm of space-time context, so that it is adaptive for general video data.

We have carried out extensive experiments on sports, news, and home video data. Compared with previous works, our approach is able to characterize the moving objects in their natural shapes and boundaries. And the processing speed on an eight-core machine is about 51.1 frames/second. The rest of this paper is organized as follows: Section 2 presents the system flow and key modules. Then Section 3 presents the workload analysis and Section 4 shows the experimental results. Last Section 5 concludes the paper and future work.

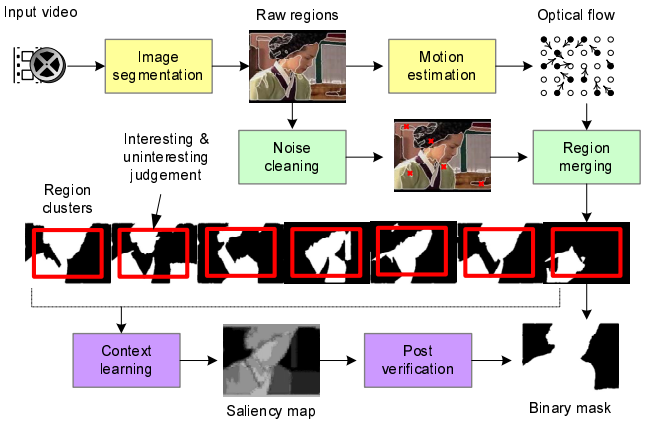

Fig. 1. System flow of the proposed detection approach.

2. SYSTEM DESIGN AND MODULE IMPLEMENTATION

2.1. Overall system

The task of automated detecting salient object is challenging as we do not know how many objects exist and how they look like for a general case. From the users' perspective, salient object in video data usually describes a set of regions whose appearances in color, texture, and motion are quite distinctive from either background or other foreground. Based on this observation, we propose a fine-grained detection approach which consists of three main phases: (I) extracting raw regions, (II) identifying interesting regions, and (III) calculating motion saliency. Figure 1 shows the system flow of the proposed approach.

The primary phase is to find out which pixels belong to an object. An object may consist of several parts whose visual appearances differ from each other. Therefore we first find out which pixels belong to a color-texture homogenous region. Shown in Figure 1, the input video frame is segmented into tens of regions, which are called super-pixel as they have closer relations with the object. An object is usually over-segmented if only considering the color-texture features. Compared with the background, regions that belong to one object often have consistent motion direction and intensity. Therefore, we extract optical flow to capture the motion characteristics and to find out which raw regions belong to one object.

The second phase is to group the raw regions in terms of motion into two basic clusters: background and foreground. Raw regions that are segmented from adjacent frames may be inconsistent due to illumination, color quantization, etc. To reduce noise disturbance, we compare the image segmentation results for every two adjacent frames and correct the inconsistent raw regions. Given the noise cleaning results, we still do not know how many salient moving regions exist. It is difficult to assume a proper distribution model to represent their characteristics. Therefore a non-parametric clustering algorithm is more suitable than a parametric clustering algorithm for our case. Shown in Figure 1, we simply assume that background is uninteresting (in black) and foreground is interesting (in white). And the red rectangle in Figure 1 helps to tell foreground from background.

The third phase is to verify the salient object by learning the space-time context. In the above region merging module, only the motion features of two adjacent frames (space context) are considered. And the distinction results of interesting and uninteresting regions are not stable for consecutive video frames. Because motion characteristics in video are sequential and smooth, the unstable results would be complementary for each other. Shown in Figure 1, by considering the consecutive frames (time context), we may learn the statistics of the interesting and uninteresting results respectively. In the probabilistic template, the lighter the element is, the higher possibility it belongs to the salient object. Such a gray-level template is an object-level saliency map. It can be used to predict the binary mask of salient object for each frame.

2.2. Key modules

2.2.1. Extracting raw regions

There have been great efforts on the task of image segmentation. Martin [7] has tested them on Berkeley Segmentation Dataset. In our experiments, the EGBIS algorithm gives better segmentation accuracy and processing speed over various testing image data. Two free parameters: σ smooths the input image before segmenting it, and k defines the segmentation threshold. We use the default values: $\sigma = 0.5$ and $k = 500$ in our experiments. Given an input image of size $W \times H$, the output result of image segmentation module is represented with a mask $I_R = [r_{ij}]_{W \times H}$, where $r_{ij} \in \{1, \cdots, R\}$, and R denotes the number of segmented raw regions.

Optical flow is the velocity field which warps one image into another (usually very similar) image. Intel Open Source Computer Vision Library provides four kinds of optical flow estimation methods. In our experiments, block matching based method is more robust than the other three. We quantize the optical flow into 9 bins according to its magnitude and direction. For each raw region, we calculate a 9-dimensional histogram to characterize its motion features: $F_m = [h_1, \cdots, h_9]$, where $h_i = n_i / \Sigma_j \{n_j\}$, n_j denotes the number of optical flow in the j^{th} bin. Evaluations show that block size of 8 could provide a robust performance.

2.2.2. Identifying interesting regions

To group raw regions into clusters for general video data, the non-parametric clustering algorithm is more appropriate than parametric modeling algorithm. In the machine learning community, spectral clustering [8] is a well received non-parametric clustering algorithm. It is crucial to define ap-

Table 1. Machine specification in workload analysis.

CPU	3.20 GHz	L1 cache	16 KB
RAM	2.0 GB	L2 cache	2048 KB
External FSB	6.4 Gb/s	L3 cache	8 MB

propriate weight of edges. Traditional definition is: $w_{ij} = \exp(-d(v_i, v_j)^2/2\sigma^2)$, where $d(v_i, v_j)$ denotes the distance between two data points v_i and v_j, and σ denotes a scaling factor. Considering that the sparse and dense clusters may co-exist, a uniform scaling factor may not work well for all the data points. Therefore we propose a local scaling factor σ_i for each local data group. Intuitively, for a sparse data group, scaling factor σ_i rapidly falls off with the distance; otherwise, σ_i slowly falls off. The local scaling factor is then defined as $\sigma_i = \sum_{j|v_j \in \Omega(v_i)} d(v_i, v_j)/n_b$, where $\Omega(v_i)$ denotes the n_b nearest neighborhoods of v_i. Accordingly, the weight of edges is defined as $w_{ij} = \exp(-d(v_i, v_j)^2/2\sigma_i\sigma_j)$. In our experiments, $n_b = 5$ works well over various video data.

2.2.3. Calculating motion saliency

To learn the statistics in space-time context, for each pixel (x, y), we calculate the element of probabilistic matrix as $P(x, y) = \Sigma_{l=1}^{N} \delta_l(x, y)/N$, given the number of consecutive frames N, where $\delta_l(x, y) = 1$ if (x, y) belongs to the interesting region cluster at frame l, otherwise $\delta_l(x, y) = 0$. After going through all the pixels within these video frames, we get the statistics in space-time context as shown in Figure 1. In order to verify the unsatisfying region merging result, for a region r given frame l, we calculate its conditional possibility of belonging to the salient moving object as $P_l(r) = \Sigma_{m=1}^{M} P_l(x_m^{(r)}, y_m^{(r)})$, where M denotes the number of total pixels within region r.

When we verify the salient objects with the context learning result, we propose an adaptive criterion to identify both the slow and fast movements. The post-verification is: r belongs to the interesting region, if $P_l(r) \geq \sqrt{S_r/S}$, where S_r denotes the size of region r, and $S = W \times H$ denotes the size of frame l. The idea behind this criterion is: a smaller-sized region is more sensitive to the motion change between adjacent frames, and thus a lower threshold is more reasonable to identify the salient movement. On the contrary, a larger-sized region is less sensitive to the motion change between adjacent frames, a higher threshold is more reasonable.

3. WORKLOAD ANALYSIS

The workload analysis has been developed on Intel Xeon MP "Tulsa" machine, whose specification is given at Table 1. With the Intel VTune Performance Analyzer [9], we profile the key modules and identify the hot spots for the proposed approach. The most two expensive modules are image segmentation (64%) and motion estimation (32%). Our primary task before parallelization is to optimize them at fine-grained level. We have optimized their algorithm flow, data structure, and calling functions. Figure 2 shows the processing speed on CIF-size image. The original approach (O_1) is 0.7 fps (frames per second), after optimizing image segmentation module (O_2) it is 3.0 fps, and after optimizing motion estimation module (1T) it is 7.7 fps.

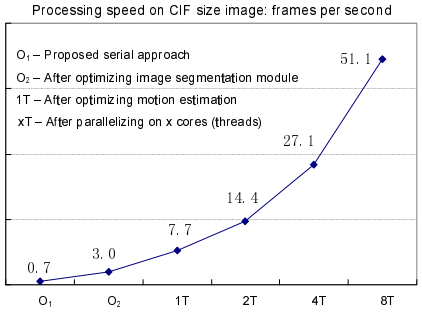

Fig. 2. Processing speed of the original approach, after module optimization, and on multiple cores (threads).

After optimizing the hot spots in serial version, we parallelized the whole algorithm over multi-core architectures. OpenMP [10] is used for the multi-threaded and shared memory parallelization. Relative to the result of 1T, the speed up is 1.87x (2-core), 3.52x (4-core), and 6.64x (8-core) respectively. It achieves a near linear speed up along with the increasing number of cores. Currently, the processing rate on an eight-core machine is 51.1 fps.

4. EXPERIMENTS

We have demonstrated the proposed approach on sports, news, and home video data. Figure 3 shows two examples, where (a) is from 2006 FIFA World Cup, and (b) is from TRECVID Rush 2007 data. The advertising board in (a) and the waving surface in (b) make it difficult to detect the salient moving objects: referee and boat. The first row in Figure 3 shows the raw region extraction results, where inconsistent regions exist for most adjacent frames. The second row shows the consistent regions after noise cleaning, where some over segmentation has been corrected. The third row shows the interesting / uninteresting distinction results, where only the motion features between adjacent frames are considered. At the end of the third row, the learning result of space-time context is given where lighter color means higher possibility of being salient movements. The last row shows the output salient object mask, where the fine-grained shape and boundary have been identified for every frame.

Figure 4 shows the comparison results with related works [1, 2, 5]. Some drawbacks are visible if we intend to use them in the advanced applications such as action recognition. For instance, the pixel-level result [1] is insufficient to tell the object appearance, and the grid-level result [5] is too coarse to

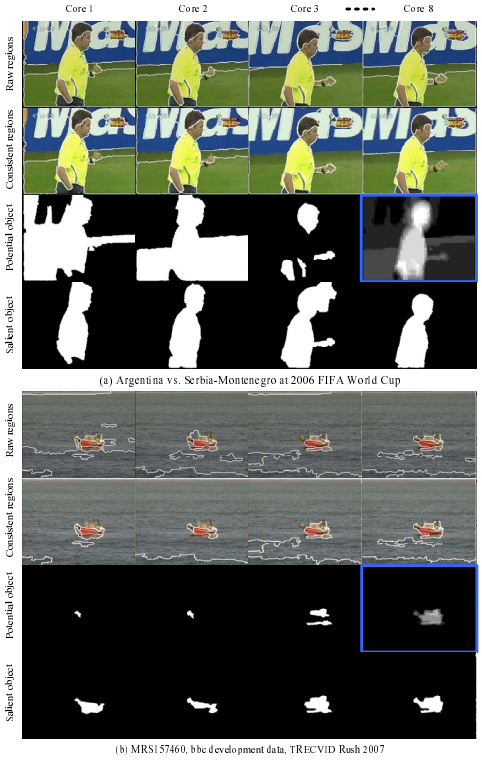

Fig. 3. Detection results on sports video and news video.

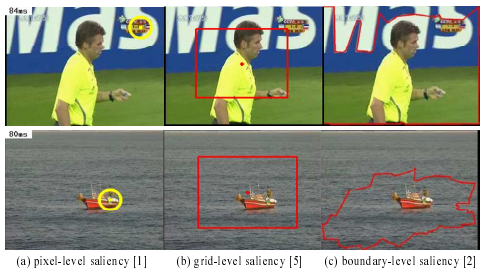

Fig. 4. Comparison results with related works.

tell the object shape. Our approach is more accurate than the boundary-level result [2], as we learn the space-time context other than just the space context. Our approach is able to tell the position and shape of salient moving object for general video data.

Figure 5 shows more experimental results of our approach. The global view of sports video, middle view of home video and close up of news video are listed from the left to the right. Both of the binary object mask and gray-level map have been detected. Generally, our approach is more robust to deal with the middle view and close up. For global view, the background may disturb the performance of "*Identifying interesting regions*". For instance, the pavilion in (a) introduces lots of inaccurate optical flow by the "*motion estimation*" module.

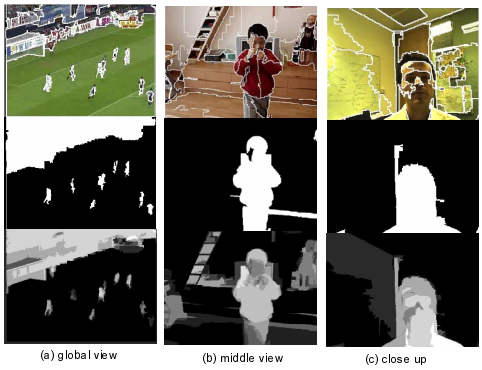

Fig. 5. Detection results on various view types.

5. CONCLUSIONS

This paper proposed a scalable parallel approach to salient moving object detection. To characterize the natural shape of salient object, we propose a bottom-up approach with unsupervised learning of space-time context. Through optimization and parallelization, the processing speed of our approach achieves 51.1 fps on the CIF size image. Extensive experimental results on various video data demonstrate the efficacy of the proposed approach. It is able to provide accurate and applicable mid-level feature for advanced applications. It indicates that the compute intensive algorithms in the multimedia field have opportunity to enhance their performance in parallelism.

6. REFERENCES

[1] L.Itti, C. Koch, and et al., "A model of saliency-based visual attention for rapid scene analysis," *PAMI*, 1998.

[2] S. Wang, T. Kubota, and J. M. Siskind, "Salient boundary detection using ratio contour," *NIPS*, 2003.

[3] Y. Gao and J. Fan, "Automatic function selection for large scale salient object detection," *ACM Multimedia*.

[4] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," *IEEE CVPR*, 2007.

[5] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," *ACM Multimedia*, 2006.

[6] C.-T. Chu, S. K. Kim, and et al., "Map-reduce for machine learning on multicore," *NIPS*, 2007.

[7] V. Martin, N. Maillot, and M. Thonnat, "http://www-sop.inria.fr/orion/personnel/vincent.martin/segmentation/segmentation.html," .

[8] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *NIPS*, 2001.

[9] Intel Vtune Performance Analyers, "http://www.intel.com/software/products/vtune/," .

[10] "Openmp application program interface," *Version 2.5*, May 2005.