Cascade Framework for Object Extraction in Image Sequences

Peng Li[1], Zhipeng Cai[2], Cheng Wang[2,3*], Zhuo Sun[2], Hanyun Wang[1], Jonathan Li[3,4]

[1]School of Electronics Science and Engineering
National University of Defense Technology
Changsha, China
*E-mail: cwang@xmu.edu.cn
[2]RSSI Lab, Department of Computer Science
Xiamen University
Xiamen, China

[3]School of Information Science and Engineering
Xiamen University
Xiamen, China
[4]GeoSTARS Lab, Faculty of Environment
University of Waterloo
Waterloo, Canada
E-mail: junli@uwaterloo.ca

Abstract— **This paper proposes a novel cascade framework to improve spatiotemporal object extraction algorithms for unconstrained image sequences. The cascade framework successively incorporates the constraints on the size of objects for candidate region prediction, an improved backprojection algorithm for coarse localization, ASIFT feature matching for object markers propagation and a novel interactive region merging method for the exact object contour segmentation. Real-world experiments show the effectiveness of the proposed method in the case of varying viewpoint, changing backgrounds, and similar distractors.**

Keywords- object extraction; localization; segmentation; backprojection; ASIFT;

I. INTRODUCTION (HEADING 1)

Recently, object extraction in image sequences has become a new challenging research direction and been applied in content-based retrieval, video coding (MPEG-4, MPEG-7), surveillance, video synthesizing, etc. [1, 2].

The object extraction approaches interpret any image/video sequence content and extract semantically meaningful objects automatically or semi-automatically by dividing video frames into patches that may correspond to objects in the scene [3, 4]. With the advance of content-based multimedia applications, a large number of automatic and semiautomatic object extraction methods have been proposed [1-4].

The state-of-the-art object extraction approaches can be divided into two categories [5]:

Category 1: The 3D volume-based approaches [6-8] construct a spatiotemporal volume by stacking consecutive frames along the temporal dimension and extract the object from the volume directly. The general scheme of the 3D volume-based methods is first to oversegment the spatiotemporal of the sequence volume into voxels and then cluster the voxels on prior knowledge of the object. The use of this categorical method is limited by the computationally expensive oversegmentation preprocessing and requirement for user interaction.

Category 2: The 2D frame-by-frame based approaches [5, 9, 10] extract objects in image/video sequences frame by frame. In general, pixels in a frame are labeled foreground or background according to color, intensity, texture, temporal consistency, motion or a mixture of them. Extraction results of the previous one or several frames provide prior information of motion, shape, color of the OOI, as well as the temporal consistency inter several adjacent frames. Compared to the 3D volume-based method, the 2D frame-by-frame based methods require less interaction and are more practical for consideration of feasibility.

Many vision techniques, including object tracking, image segmentation, clustering, etc. are involved in the task of image sequence object extraction [5]. Recently, Stalder et al. [11] proposed a cascaded confidence filtering (CCF) for object tracking. The CCF cascaded filters of different vision cues, i.e. geometry, background statistics, temporal consistency, to reduce the searching region objects. Promising results demonstrated the power of cascaded filters. However, the CCF approach has the constraint of static cameras due to the use of background statistics. Moreover, it can only extract the object candidate region rather than the exact object contour.

In this paper, we propose a novel cascaded object extraction scheme for unconstrained image sequences by combine the 2D frame-by-frame with the idea of the CCF approach [11]. In contrast to the CCF approach, the input of our method is the original image rather than a confidence score obtained by some detectors. We improve the classic histogram backprojection method [12] for coarse localization of the objects by combining the object and background information together. Geometric and local invariant features are introduced to assist the localization approach. An interactive marker based region merging algorithm [13] is introduced to extract the exact contour, which cannot be obtained by the CCF approach. The markers are specified by the users for the first keyframe and propagated by the ASIFT [14] automatically for the subsequent frames.

This remainder of this paper is organized as follows: Section II shows the details of the proposed algorithm. Section

This paper is supported by National Natural Science Foundation of China Project 40971245.

III presents experimental results and Section IV summarizes the paper.

II. THE PROPOSED METHOD

A. Overview

The proposed approach is elaborated to extract object of interest in image sequences and can be used in many other similar problems. The inputs of the approach include an initialization of the object of interest in a key frame and the rest sequence frames. In this paper, the initialization of the object is indicated by the user with a bounding box. Nevertheless, the initialization can be given by any other object detectors. As shown in the Fig. 1, there are following four cascaded steps for each frame. First, according to the geometric information of the object in the previous frame and the temporal consistency of the image sequence, we detect the candidate region of the object in the current frame. Second, we coarsely localize the object by a new color histogram backprojection algorithm, which takes both the background and object information into account simultaneously. Third, we refine the coarse localization result by local feature matching. Then the object is accurately localized by a bounding box. In this paper, we select the recently proposed ASIFT (affine scale invariant feature transform) [14]. Finally, the exact object contour is segmented by maximal similarity based region merging algorithm (MSRM) [13]. Then the object segmentation will be feedback for the approach. Graphical explanation of each step is show in Fig. 4.

B. Candidate Region Prediction

The position change of the object through consecutive frames is described by a second-order autoregression (AR) model, i.e.:

$$p_n = Ap_n - 1 + Bp_{n-2}, \tag{1}$$

where state p denotes the position of the object, A and B are the model parameters, which can be learned from the previous tracks. When the object's movement is smooth, AR model can generally provide better prediction result. However, object in unconstrained sequences usually changes the direction with sudden movement, which leads to the bad prediction for the following steps. In order to avoid this worse case, the candidate region is set much larger than the object. Usually, both the width and the height of the candidate region are set 2 times of the object bounding box's. However, in most cases, the candidate region is still much smaller than the entire frame.

C. Coarse Localization

To avoid exhaustive searching in the whole image, we propose an improved histogram backprojection was proposed to handle the coarse location problem efficiently.

Color histogram back-projection is a low complexity, efficient vision algorithm for finding objects in complex scenes and is little affected by the movement of camera. Assuming the color quantized into N indexes, the histogram backprojection method computes the ratio $\{R_i\}_{\mathrm{i}}$ between the color histogram of the object of interest $\{O_i\}_{\mathrm{i}}$ and the color histogram of the entire image $\{I_i\}_i$, as shown in (2):

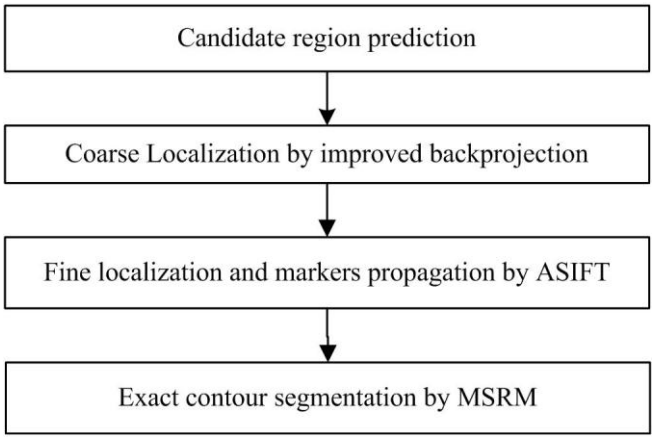

Figure 1. Flowchart of the cascaded processing steps.

$$R_i = O_i / I_i \quad i = \{1, ..., N\}. \tag{2}$$

R_i can be explained as the posterior of the occurrence of the OOI when observing pixel indexed i [12], i.e., $P(O|i) = P(i|O)P(O)/P(i)$. A probability image, I_{op}, similar to the confidence score for the CCF approach, is obtained by calculating the posterior for all the pixels in the image.

From (2), it can be seen that the classical only take the foreground information into account and the background information has been abandoned. However, for the tracking applications, the background near the foreground OOI is less changeable due to the temporal consistency of sequence and meaningful for occurrence of the OOI.

To make full use of the near background of the OOI, we compute the color histogram of the near background $\{B_i\}_{\mathrm{i}}$. Then we backproject $\{B_i\}_{\mathrm{I}}$ to the test image and get the near background probability image I_{bp}. Then the new probability image is calculated as a weighted sum of $\mathrm{I_{op}}$ and $\mathrm{I_{bp}}$, i.e.,

$$I_{OP} = \lambda_1 I_{op} + \lambda_2 I_{bp}, \tag{3}$$

where λ_1 and λ_2 are two positives summing to 1. By convolving the new probability image with a disk filter, the object and near background contribute to the occurrence probability image, I_{COP}. Then the object is coarsely localized at the local maxima in the candidate region and the searching region is further reduced. An example of the improved backprojection approach is shown in Fig.2. Compared with the classic backprojection approach, the detected candidate regions are more accurate.

In contrast to the mixture model based methods [11], the proposed method does not rely on the assumption of static background and require less computation.

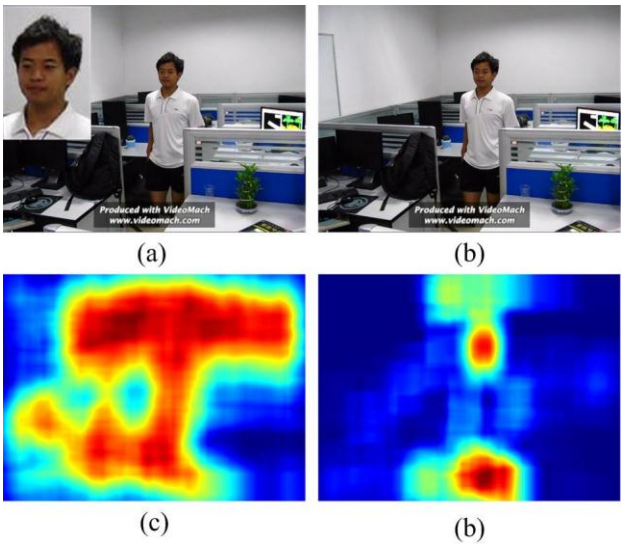

Figure 2. Demonstration of the improved backprojection algorithm. (a) The 8th frame and the face of interest. (b) The 9th frame. (c) Confidence score of the classic backprojection algorithm. (d) Confidence score of the proposed backprojection algorithm.

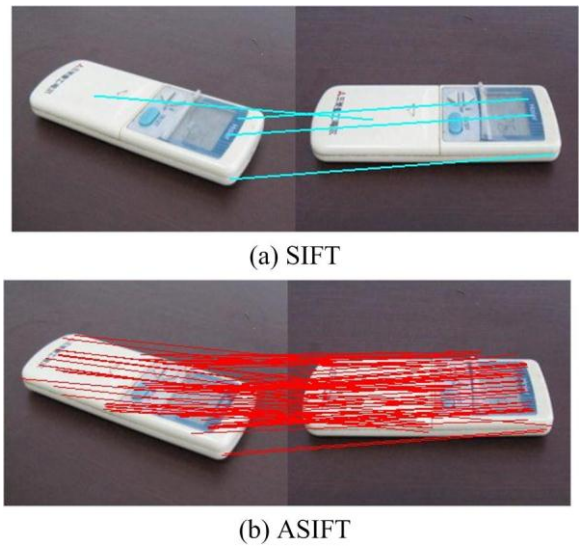

Figure 3. Comparison of SIFT and ASIFT on a pair of images of a remote control. (a) There are only 4 matched SIFT features. (b) There are 89 matched SIFT features.

D. Fine Localization and Marker Propagation

The backprojection results give a coarse localization of the OOI. To refine the localization results, we apply the ASIFT to localize and find the affine transform for the OOI and the markers for segmentation.

ASIFT is an affine invariant extension of SIFT that extracts keypoints from the image and creates a high dimensional description vector (descriptor) for the local image content [14]. Compared to classic SIFT, ASIFT fully affine invariant by simulating the two camera parameters and then applying SIFT which simulates the scale and normalizes the rotation and the translation. Therefore, the ASIFT can produce much richer features than the classic SIFT, as demonstrated in Fig. 3.

When refining the localization results, first, we calculate ASIFT features in both the OOI template and the candidate region; Second, we find matching features and next employ the optimized random sample algorithm (ORSA) method to eliminate the outliers and calculate the affine transformation T of the OOI. We translate the object template image to the candidate region with T. The bounding box of the translated template is the fine localized candidate region.

E. Object Contour Segmentation

To get the exact contour of the OOI, we adopt the recently proposed maximal similarity based region merging (MSRM) method [13], which is an interactive method based on the initial segmentation of mean shift. The interactive information is introduced as markers, i.e. M, which are inputs by the users to roughly indicate the position and main features of the object and background. The key contribution of the method is a novel maximal similarity based region merging (MSRM) mechanism, which is adaptive to image content and does not require a preset threshold. With this algorithm, the non-marker background regions will be automatically merged and labeled, while the non-marker object regions will be identified and avoided from being merged with background.

For the initial keyframe of the entire sequence, the interactive markers M are manually labeled by the users. For the subsequent frames, the new markers M' are obtained by transform M according to the transformation matrix of the OOI, i.e.,

$$M' = TM \,. \tag{4}$$

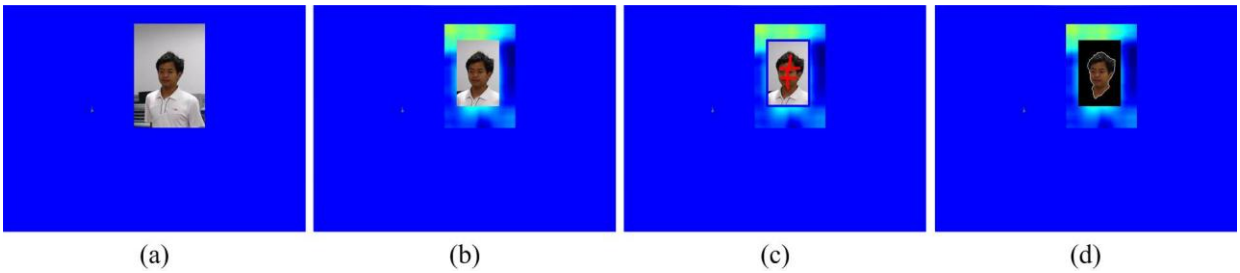

Figure 4. Results of each steps of the entire extraction approach. (a) Candidate region predication. (b) Coarse localization by improved backprojection. (c) Marker propagation and fine localization by ASIFT. (d) Object contour extraction.

Figure 5. Face extraction from a video sequence. The contour of the face is shown in red.

III. EXPERIMENTAL RESULTS

In this section, we present the experiment results of the proposed cascaded object extraction approach. Figure 4 shows the result of each steps of the entire procedure.

In Fig. 5, we show the extraction of a face in a video sequence captured by a handheld camera.

To further validate the proposed approach, we apply the approach to digital measurable image (DMI) sequences captured by the VISAT^{TM} mobile mapping system. The frame rate of the sequence is 1 frame per second and the object are undergoing large translation and viewpoint change. The experimental results of several different GIS OOI are shown in Fig. 6.

The above experimental results demonstrate that the propose approach extend the CCF approach [11] successfully and can extract objects from image sequences effectively.

IV. CONCLUSIONS

A cascade framework for is proposed for spatiotemporally extracting objects in image sequences. The cascaded framework involves geometric constraints for candidate region predication, an improved backprojection method for coarse location, ASIFT matching for marker propagation and MSRM for contour extraction. The experimental evaluation on real image sequences demonstrates the effectiveness of the proposed method. Compared to the CCF method [11], the proposed framework doesn't rely on the static background hypothesis and can obtain the exact object contour.

ACKNOWLEDGMENT

We appreciate Qiugen Xiao with Xiamen University greatly for his support with our experiments.

Figure 6. GIS OOI extraction for DMI sequences. The localization bounding boxes of the OOIs are shown in magenta and the contours are shown in yellow.

REFERENCES

[1] F. Porikli, F. Bashir, and S. Huifang, "Compressed Domain Video Object Segmentation," IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, no. 1, pp. 2-14, January 2010.

[2] C.-Y. Chung, and H. H. Chen, "Video Object Extraction via MRF-Based Contour Tracking," IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, no. 1, pp. 149-155, January 2010.

[3] J. C. Nascimento, and J. Marques, "Performance evaluation of object detection algorithms for video surveillance," IEEE Transactions on Multimedia, vol. 8, no. 4, pp. 761-774, Augest 2006.

[4] P. L. Correia, and F. Pereira, "Classification of Video Segmentation Application Scenarios," IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 5, pp. 7, May 2004.

[5] P. Li, and C. Wang, "Object of interest extraction in low-frame-rate image sequences and application," Optical Engineering, vol. 51, no. 6, pp. 06720101-06720112, June 2012.

[6] M. Grundmann, V. Kwatra, M. Han et al., "Efficient hierarchical graph-based video segmentation." CVPR, 09

[7] J. Wang, P. Bhat, R. A. Colburn et al., "Interactive video cutout," Acm Transactions on Graphics, vol. 24, no. 3, pp. 585-594, July 2005.

[8] Y. Li, J. Sun, and H. Y. Shum, "Video object cut and paste," Acm Transactions on Graphics, vol. 24, no. 3, pp. 595-600, July 2005.

[9] J. Lezama, K. Alahari, J. Sivic et al., "Track to the future: Spatio-temporal video segmentation with long-range motion cues." CVPR, 2011

[10] X. Bai, J. Wang, D. Simons et al., "Video SnapCut: Robust Video Object Cutout Using Localized Classifiers," Acm Transactions on Graphics, vol. 28, no. 3, Aug, 2009.

[11] S. Stalder, H. Grabner, and L. Van Gool, "Cascaded Confidence Filtering for Improved Tracking-by-Detection," ECCV, 2010.

[12] M. J. Swain, and D. H. Ballard, "Color indexing," International Journal of Computer Vision, vol. 7, no. 1, pp. 11-32, Nov, 1991.

[13] J. F. Ning, L. Zhang, D. Zhang et al., "Interactive image segmentation by maximal similarity based region merging," Pattern Recognition, vol. 43, no. 2, pp. 445-456, Feb, 2010.

[14] G. S. Yu, and J. M. Morel, "A fully affine invariant image comparision method," International Conference on Acoustics Speech and Signal Processing, 2009.