Journal of Shanghai University (English Edition), 2004, 8(1): 70~74
Article ID: 1007-6417(2004)01-0070-05

Objective Performance Evaluation of Video Segmentation Algorithms with Ground-Truth

YANG Gao-bo(杨高波), *ZHANG Zhao-yang*(张兆扬)
Schol of Communication and Information Engineering, Shanghai University, Shanghai 200072, *P.R. China*

Abstract While the development of particular video segmentation algorithms has attracted considerable research interest, relatively little effort has been devoted to provide a methodology for evaluating their performance. In this paper, we propose a methodology to objectively evaluate video segmentation algorithm with ground-truth, which is based on computing the deviation of segmentation results from the reference segmentation. Four different metrics based on classification pixels, edges, relative foreground area and relative position respectively are combined to address the spatial accuracy. Temporal coherency is evaluated by utilizing the difference of spatial accuracy between successive frames. The experimental results show the feasibility of our approach. Moreover, it is computationally more efficient than previous methods. It can be applied to provide an offline ranking among different segmentation algorithms and to optimally set the parameters for a given algorithm.

Key words video object segmentation, performance evaluation, MPEG-4.

1 Introduction

MPEG-4 introduces the concept of video object to support content-based functionalities. However, it is not specified in the standard how the video objects are generated. So video segmentation is a crucial factor in the future success of MPEG-4. Many video segmentation algorithms have been proposed in the literature. These algorithms use different sets of techniques, and result in different performance. Most of them are application-dependent, *i.e.*, their performance varies with different image sequences. However, relatively little effort has been made to evaluate the performance of video segmentation algorithms. A widely accepted methodology to evaluate them is still absent. While in image coding and computer vision literature, the most frequently used metric to evaluate deviations between the original and the coded images is peak-signal-to-noise ratio (PSNR), a widely accepted metric is also needed to objectively evaluate video object segmentation algorithms.

Received Apr. 14, 2003; Revised May 23, 2003
Project supported by the National Nature Science Foundation of China (Grant No. 60172020)
YANG Gao-bo, Ph.D. Candidate, E-mail: gbyang@sohu.com
ZHANG Zhao-yang, Ph.D. Prof., E-mail: zhyzhang@yc.shu.edu.cn

- Performance evaluation let researchers know the strength and weakness of a particular approach and promote new developments by effectively taking strong points of different algorithms. Successful evaluation provides effective guidelines in choosing a suitable algorithm according to applications, and help appropriately setting its parameters to achieve the best performance.
- In automatic video segmentation algorithm, feedback based on performance evaluation is often introduced to terminate the iteration process when no satisfactory results are obtained[1].

According to the presence or absence of reference object masks, performance evaluation of video segmentation algorithms can be classified into relative evaluation and standalone evaluation. When the reference object masks, *i.e.*, ground-truth, exist, a more accurate and robust evaluation can be achieved. So we focus on objective performance evaluation with ground-truth in this paper.

The paper is organized as follows. In Section 2, we give a brief review of the current state of research in this topic. In Section 3, we introduce and propose some novel and efficient measures for evaluation of the spatial accuracy and temporal coherency. Experimental results are given in Section 4. And we conclude this paper in Section 5.

2 State of the Art: a Brief Review

The main difficulty in addressing video segmentation and its performance evaluation stems from the ill-defined nature of the problem itself. As such, it is difficult to define a universally "good" segmentation. Since human visual system (HVS) can identify and interpret scenes with different semantic objects effortlessly, and human observers are the end users in multimedia applications, metrics in accordance with HVS seem to be more appropriate in predicting user acceptance. In general, two types of analysis must be made when evaluating the segmentation algorithms[2]:

- Spatial accuracy: how close the segmented object masks resemble the reference masks in every frame;
- Temporal coherency: stability and evolution of the estimated masks along time.

Because the lack of temporal stability of segmentation masks is often considered as one of the most annoying artifacts when viewing an entire sequence, temporal coherency is more important than spatial accuracy in video segmentation. However, temporal coherency is the extent to which every segmented mask resembles the reference masks, *i.e.*, the spatial accuracy of every frame. So spatial accuracy plays a fundamental role and is crucial for video segmentation.

In the literature, the first approach for objective evaluation of VOP generation algorithms was motivated by the core-experiment of MPEG-4[3]. The spatial accuracy of an estimated object mask in frame t is defined as

$$d(A_t^{\text{est}}, A_t^{\text{ref}}) = 1 - \frac{\sum_{(x,y)} A_t^{\text{est}}(x,y) \oplus A_t^{\text{ref}}(x,y)}{\sum_{(x,y)} A_t^{\text{ref}}(x,y)},$$

where A_t^{est} and A_t^{ref} are the reference and the estimated object masks in frame t respectively, and $\oplus$ is the logical "XOR" operation. This approach is rather simple. However, the role of HVS is not incorporated in this approach.

Villegas[2] utilizes a number of properties that make some errors visually more important than others, and classifies two types of errors, namely missing foreground points and adding background points, and then tries to weight the quality measure values so as to take them into consideration. Similarly, Erdem[4,7] proposes three evaluation metrics based on weighted misclassification penalty, shape penalty and motion penalty. Correia[5,6] utilizes some statistical data similarity and geometrical similarity such as compactness, elongation to describe the spatial accuracy. Cavallaro[8] proposes a methodology based on computing the deviation of the segmentation results from reference segmentation. The discrepancy is weighted based on spatial and temporal contextual information. Most of the above approaches incorporate the role of HVS. However, they suffer from a common drawback: high computational complexity. In addition, some concepts such as compactness, elongation and circularity in these metrics are obscure and ill-defined.

COST 211 launched a campaign for comparing video segmentation algorithms to the COST 211 quat analysis model (AM). An exchange platform for algorithms and sequences related to video segmentation was established. For objective evaluation of the segmentation results, the three criteria proposed in Ref. [3] were utilized. For further information, please refer to its website[9].

3 Proposed Method

From the discussion above, we can conclude that a good method for performance evaluation should meet the following two requirements: (1) To incorporate the role of HVS; (2) Easy interpretation and computational simplicity. Since the mechanism of HVS is still hard to be modeled at present, the most straightforward method is to take into account those factors that are sensitive to HVS, and to weigh them according to their visual relevance.

3.1 Spatial accuracy

3.1.1 Pixel classification based measure (PCM)

Video object segmentation can be seen as a classification problem. It describes a pixel in the original image to be classified as a foreground pixel or background pixel, and then masks all the foreground pixels. Pixel classification based measure (PCM) is common and straightforward. It reflects the percentage of background pixels wrongly assigned to the foreground and conversely foreground pixels wrongly assigned to the background. It is simply defined as

$$PCM = 1 - \frac{Cardi(B_{\text{ref}} \cap F_{\text{seg}}) + Cardi(F_{\text{ref}} \cap B_{\text{seg}})}{Cardi(F_{\text{ref}}) + Cardi(B_{\text{ref}})},$$

where B_{ref} and F_{ref} denote the background and foreground of the reference image (ground-truth), while

B_{seg} and F_{seg} denote the background and foreground area pixels in the segmentation result. "$\cap$" is the logical "AND" operation. Cardi($\cdot$) is the cardinality operator. Obviously, *PCM* varies from zero for a totally wrongly segmented image to 1 for a perfectly classified image.

3.1.2 Edge-based measure (EM)

In the perception of scene content by HVS, edges play a major role. Similarly, machine vision algorithms often rely on feature maps obtained from the edges. Thus task performance in vision, whether by human or machine, is highly dependent on the quality of the edges. We adopt an edge-based measure by Pratt, which considers both edge location accuracy and missing/false edge elements[10]. The measure is based on a priori ideal reference edge map. The figure of the merit is defined as

$$EM = \frac{1}{\max(n_{seg}, n_{ref})} \sum_{i=1}^{n_{seg}} \frac{1}{1 + \alpha d_i^2},$$

where n_{seg}, n_{ref} are the number of edge points of the segmented image and ideal reference segmentation respectively. And d_i is the distance to the closest edge pixel for the i-th detected edge pixel of the segmented video object. The factor $\max(n_{seg}, n_{ref})$ penalizes the number of false alarm edges or conversely missing edges ($\alpha > 0$, often $\alpha = 1/9$). The scaling factor α provides a relative weighting between smeared edges and thin but offset edges. The sum terms penalize possible shifts from the correct edge positions. For a correctly segmentation, EM will be 1 and decrease for increasing edge discrepancy.

3.1.3 Relative foreground area measure (RFAM)

Good image segmentation should yield accurate measurements on the object properties such as area and shape. The comparison of these measurements obtained from the segmented image with respect to the reference image provides useful discrepancy measures. Here we propose a relative foreground area measure (RFAM). And it is defined as

$$RFAM = 1 - \frac{|Area(I_{ref}) - Area(I_{seg})|}{Area(I_{seg})}.$$

Obviously for a perfect match of the segmented regions *RFAM* is 1. In general, most video segmentation algorithms can get accurate results, so the difference between the segmented object area and the reference object area is small. *RFAM* will be in the range of [0,1].

3.1.4 Relative position based measure (RPM)

RPM is defined as the centroid shift between reference and segmented object masks. It is normalized by the perimeter of the reference object (assume a circular objectis), so *RPM* can be expressed as

$$RPM = 1 - \frac{\| Cent_{ref} - Cent_{seg} \|}{2\sqrt{\pi} \times \sqrt{\text{area}_{ref}}}.$$

Here $\| \cdot \|$ is the euclidean distance. The centroid of reference object can be expressed as

$$Cent_{ref}(x) = \frac{\sum_{(x,y)\in I_{ref}} x}{\sum_{(x,y)\in I_{ref}} 1},$$

$$Cent_{ref}(y) = \frac{\sum_{(x,y)\in I_{ref}} y}{\sum_{(x,y)\in I_{ref}} 1}.$$

Similarly, the centroid of segmented object mask can be computed. *RPM* will be 1 for a perfect segmentation, and decrease with the centroid shift increases.

3.1.5 Combination of the afore-mentioned measures

The above four metrics *PCM*, *EM*, *RFAM* and *RPM* are addressing spatial accuracy from different aspects. A single numerical measure can be obtained to evaluate the spatial accuracy (SA) of segmentation algorithm by combing the metrics defined above as follows:

$$SA(t) = \alpha PCM + \beta EM + \gamma RFAM + \phi RPM,$$
$$(\alpha > 0, \beta > 0, \gamma > 0, \phi > 0, \alpha + \beta + \gamma + \phi = 1)$$

where the parameters α, β, γ, and ϕ are weighting coefficients according to their visual relevance to HVS. Compared with *PCM*, *EM* and *RFAM*, RPM is less sensitive to HVS, so the weighting factor of RPM is less than that of others. From our experiments, we found that $\alpha = 0.3, \beta = 0.3, \gamma = 0.3, \phi = 0.1$ is an appropriate selection. However, it can also be adjusted for on-line performance evaluation according to the characteristics of the video sequence and the relative importance and accuracy of the measures above. Because *PCM*, *EM*, *RFAM* and *RPM* are in the range of [0, 1], the combined spatial accuracy metric $SA(t)$ will takes the values between [0,1].

3.2 Temporal coherency

Temporal coherency (TC) is the extent to which every segmented mask resembles the reference mask, *i.e.*, the spatial accuracy of every frame. We define

it as the difference value of successive frames t and $t-1$:

$$TC(t) = 1 - |SA(t) - SA(t-1)| \quad (t = 1 \ldots n)$$

4 Experimentatal Results

In this section, the proposed evaluation metrics are used to compare and rank different segmentation algorithms, and to compare the performance of the same algorithm on different video sequences.

4.1 Selection of ground-truth

The ground-truth plays an important role in the evaluation. The reference segmentation can be obtained in many ways such as Chroma-keying, manual extraction of masks, or any automatic segmentation that is "good" enough. Here we use reference masks provided by COST 211[9].

4.2 Selection of video segmentation algorithms

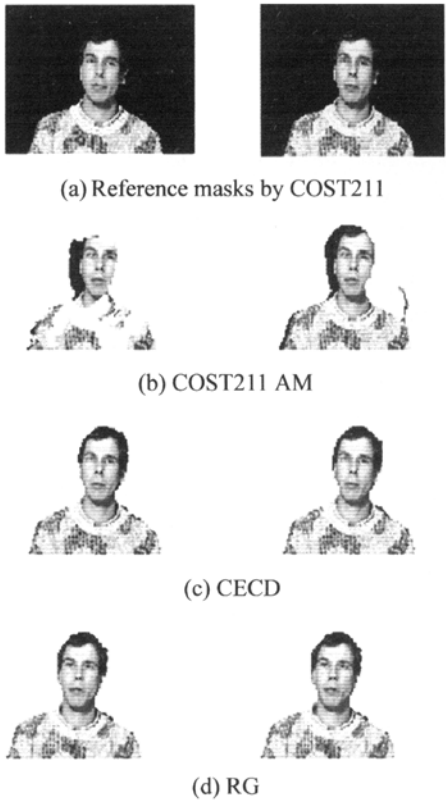
In general, the selected algorithms to be evaluated must be typical, namely, they must differ fundamentally in their algorithmic philosophy. Thus three algorithms based on COST AM[9], CECD[11] and a region growing (RG)[12] based algorithm are selected to be evaluated.

4.3 Experiment on two typical sequences

To demonstrate our approach, we choose two typical MPEG-4 sequences Erik and hall monitor from COST 211data set. There is only one video object in each frame within a simple background in Erik. Fig.1 lists some experimental results on Erik sequence.

The performance evaluation of the above three algorithms on Erik sepuence is shown in Fig.2. Fig.2 (a) is the spatial accuracy. Fig.2(b) is the temporal coherency. From the figures, we can see that RG algorithm provides the best segmentation results with a SA among [0.90, 0.97], thus its TC is about 0.96. Visually pleasant segmentation results are achieved with RG on Erik sequence. The SA of CECD is better than COST AM but its TC is worse. This is caused by the fact that video segmentation algorithm based change detection is sensitive to noise and illumination variations.

Hall Monitor sequence is a video sequence with more than one video object and a complex background. Also there are drastic changes for the video object. The SA values for Hall monitor sequences are among 0.55~0.90. Especially COST AM can achieve



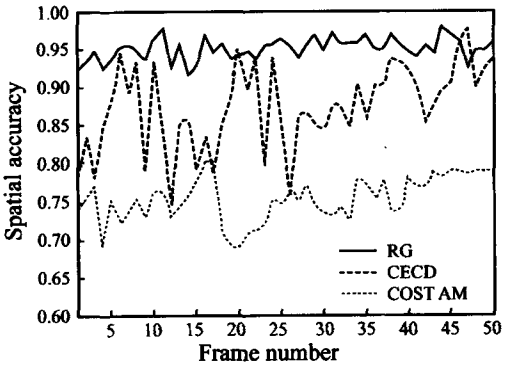
(a) Reference masks by COST211

(b) COST211 AM

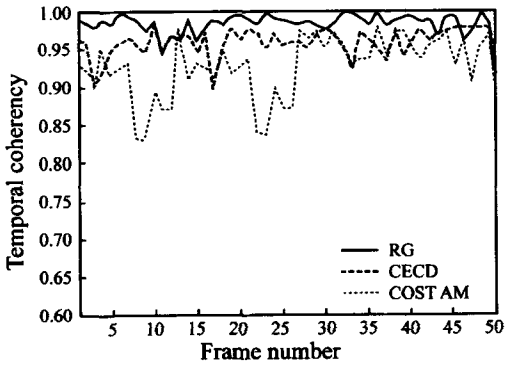
(c) CECD

(d) RG

Fig.1 Segmentation results on Erik sequence



(a) Spatial accuracy on Erik sequence



(b) Temporal coherency on Erik sequence

Fig.2 Objective comparison of segmentation on Erik sequence

the SA on Hall monitor only about 0.55 for the beginning frames. However, the TC values are still about 0.8～0.90. Due to space restriction, figures for SA and TC of the three video segmentation algorithms achieved on Hall monitor is omitted.

All the three algorithms can obtain more satisfactory results for Erik sequence than Hall monitor, which demonstrates that video object segmentation algorithms are application-dependent, and their performance varies with the contents of the video sequences.

Obviously, the experimental results of the proposed objective evaluation metrics are in accordance with subjective evaluation by human observers.

4.4 Efficiency of our approach

We also conduct experiments to compare the computational complexity of the proposed approach with respect to the approaches by Andrea[8] and Erdem[4]. It is conducted on the 1 to 10 frames of Akiyo sequence (QCIF). Experimental result is illustrated in Table 1.

The Experiment results show that the efficiency of the proposed approach is slightly improved compared with that in Ref. [4] and much more efficient than in Ref. [8] (about double processing speed). So the proposed approach is also computationally efficient.

Table 1 Comparison of execution time among different algorithms

Approach	Physical time (ms)	Computational efficiency
Erdem[4]	4.132	Acceptable
Andrea[8]	2.241	Well
Proposed approach	2.086	Well

5 Conclusion

In this paper, an objective and computationally efficient method based on spatio-temporal information is proposed to evaluate video segmentation algorithm. It incorporates the role of HVS, and the behavior is in accordance with subjective evaluation by human observers. In semiautomatic video segmentation, human interaction is an indispensable part to help provide the semantic information. However, the amount of human interaction should be as little as possible, and the user experience is different for user interaction in initial object extraction and object tracking. Future research will be concentrated on incorporating certain amount of human interaction into performance evaluation.

References

[1] Erdem C E, Sankur B, Tekalp A M. Non-rigid object tracking using performance evaluation measures as feedback [A]. *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, CVPR' 2001 [C]. Hawaii, 11-13 Dec. 2001: 323－330.

[2] Villegas P, Marichal X, *et al*. Objective evaluation of segmentation masks in video sequences [A]. *Proc. of WIAMIS 99 workshop*[C]. Berlin, May 1999: 85－88.

[3] Wollborn M, Mech R. Refined Procedure for Objective Evaluation of VOP Generation Algorithms, Doc, ISO/IEC JTC1/SC29/WG11 MPEG98/3448 [R]. Tokyo, March 1998.

[4] Erdem C E, Sandur B. Performance evaluation of segmentation masks in video sequences [A]. *EUSIPCO'* 2000: 10*th European Signal Processing Conference* [C]. Tampere, Finland, 5-8 September 2000: 917－920.

[5] Correia P, Pereira F. Objective evaluation of relative segmentation quality [A]. *Proc. of IEEE Conference on Image Processing* (ICIP2000)[C]. Canada, 10－13, Sept. 2000: 308－311.

[6] Correia P, Pereira F. Objective evaluation of standalone segmentation quality [A]. *Proc. of WIAMIS* 2001 *workshop*[C]. Tampere, 2001.

[7] Erdem C E, Tekalp A M, Sandur B. Metrics for performance evaluation of video object segmentation and tracking without ground-truth [A]. *Proc. of IEEE Conference on Image Processing* (*ICIP*2001)[C]. Thessaloniki, Greece, 2001.

[8] Cavallaro A, Drelie E, Ebrahimi T. Objective evaluation of segmentation quality using spatio-temporal context [A]. *Proc. of IEEE International Conference on Image Processing*[C]. Rochester, New York, Sept. 2002: 301－304.

[9] Working site for sequences and algorithms exchange[J/OL]. http://www.tele.ucl.ac.be/exchange

[10] Pratt W K. Digital Image Processing[M]. Wiley, New York, 1978.

[11] Cavallaro A, Ebrahhimi T. Change detection based on color edges [A]. *Proc. of IEEE International Symposium on Circuits and Systems*[C]. Sydney, Australia, 2001.

[12] Experimental results[J/OL]. http://www.csd.uoc.gr/～tziritas

(*Executive editor YAO Yue-yuan*)