New Method for Evaluation of Video Segmentation Quality

Mahmud Abdulla Mohammad, Ioannis Kaloskampis and Yulia Hicks
School of Engineering, Cardiff University, Queen's Buildings, The Parade, Cardiff, UK
{MohammadMA, KaloskampisI, HicksYA}@cardiff.ac.uk

Keywords: Image/Video Segmentation, Object Segmentation, Image Processing, Edge Detection, Statistical Measures.

Abstract: Segmentation is an important stage in image/video analysis and understanding. There are many different approaches and algorithms for image/video segmentation, hence their evaluation is also important in order to assess the quality of segmentation results. Nonetheless, so far there was little research aimed specifically at evaluation of video segmentation quality. In this article, we propose the criteria of good quality of video segmentation suitable for assessment of video segmentations by including a requirement for temporal region consistency. We also propose a new method for evaluation of video segmentation quality on the basis of the proposed criteria. The new method can be used both for supervised and unsupervised evaluation. We designed a test video set specifically for evaluation of our method and evaluated the proposed method using both this set and segmentations of real life videos. We compared our method against a state of the art supervised evaluation method. The comparison showed that our method is better at evaluation of perceptual qualities of video segmentations as well as at highlighting certain defects of video segmentations.

1 INTRODUCTION

Image segmentation means subdividing images into meaningful segments or regions (Dey et al., 2010; Morris et al., 1986). Image sequence (video) segmentation can be viewed as an expansion of single image segmentation (Charron and Hicks, 2010; Greenspan et al., 2004). The additional challenge in video segmentation is to cater for segmentation consistency throughout the video (Kaloskampis and Hicks, 2014). There are many different approaches and algorithms for image/video segmentation hence it is important to be able to evaluate the quality of the segmentations produced by different methods. To date, the evaluation methods are divided into three classes (Zhang et al., 2008; Correia and Pereira, 2003).

Subjective evaluation is the process in which human observers quantify the quality of segmentation results on the basis of visual description. This is a complicated and time consuming process, and the result varies from one observer to another.

Supervised evaluation is the process in which a segmented image/frame is compared against a manually-segmented (ground truth) reference image/frame. Producing the ground truth images is also a time consuming process, with a degree of disagreement between different people.

Unsupervised evaluation, which is also known as stand-alone evaluation or empirical goodness evaluation. It works automatically without any extra requirements such as the ground truth images. The methods in this evaluation class use only low-level features and do not incorporate semantic information.

Most of evaluation methods are subjective or related to specific applications. The majority of proposed objective evaluation methods is in the area of supervised evaluation, with the area of unsupervised evaluation receiving the least attention (Zhang et al., 2008). Evaluation is usually based on several criteria, with each criterion considering the quality of the segmentation from a different aspect. A number of researchers considered which aspects of the segmentation quality should be evaluated. In the remainder of this section, we will review the existing criteria along with metrics used to evaluate them.

For example, Levine and Nazif (Levine and Nazif, 1985) suggested that to design a measure for evaluating the quality of image segmentation, it is necessary to consider the following: *(i) uniformity within regions, (ii) contrast across regions, (iii) provision for lines and texture.* Haralick and Shapiro (Haralick and Shapiro, 1985) proposed four criteria to evaluate image segmentation: *(i) the regions must be uniform and homogeneous, (ii) adjacent regions should have significant differences with respect to the characteris-*

tic on which they are uniform, (iii) region interiors should be simple and without holes, (iv) boundaries should be smooth and accurate. Most of the previous evaluation methods and metrics incorporate the above criteria, either directly or indirectly (Levine and Nazif, 1985; Liu and Yang, 1994; Borsotti et al., 1998; Chen and Wang, 2004; Zhang et al., 2004; Chabrier et al., 2006).

Consequently (Zhang et al., 2008) classified the previous work according to the criteria proposed in (Haralick and Shapiro, 1985). The classification also covers unsupervised metrics proposed for evaluation of image and video segmentation. They concluded that these criteria had become the *de facto* standard for unsupervised evaluation of image segmentation. They suggested that the first two criteria were more characteristic rather than semantic and hence incorporated first and second criteria in their work. Zhang *et al.* (Zhang et al., 2008) conducted a comparative evaluation of different approaches after which they concluded that previous unsupervised approaches for evaluation of image segmentation methods are insufficient for comparison of segmentation produced by different algorithms. The criteria discussed above have been applied for evaluation of the quality of image segmentation. The previous unsupervised methods for evaluation of video segmentation methods (Correia and Pereira, 2003; Erdem et al., 2004) are limited and not designed for general purpose applications, with the former method manually labeling data, and the latter being designed for evaluating video object segmentation and tracking algorithms. Likewise, the metric proposed in (Gelasca and Ebrahimi, 2006) is based on spatial and temporal accuracy and designed for evaluating video object segmentation.

In addition to the above, there are several popular supervised evaluation methods based on the image/frame boundaries as opposed to regions. The boundary precision-recall is used in (Martin et al., 2001) as a supervised metric for evaluation of image segmentation. Galasso *et al.* (Galasso et al., 2013) introduced the volume precision-recall metric for evaluation of video segmentation quality, Xu *et al.* (Xu and Corso, 2012) proposed 3D volumetric quality metrics to evaluate super-voxel methods, which they based on the boundaries, without taking into account the region uniformity and consistency.

To summarise the current state of art in the area of evaluation of the video segmentation quality, the following is noted: (i) there is no established criteria for evaluation of overall video segmentation as opposed to image segmentation or video object segmentation, (ii) there is a limited number of unsupervised evaluation methods of video segmentation and they are not designed for overall video segmentation, (iii) supervised evaluation methods of video segmentation consider the boundaries of the segmentations without taking into account region interiors.

In this work, we propose new criteria of good quality of video segmentation to include temporal region consistency. On the basis of the new criteria, we propose an online method for evaluation of video segmentation quality, which takes into account the characteristics of both boundaries and regions. Online evaluation can be used to control the parameters of online video segmentation in real-time applications (Zhang et al., 2008). Proposed method can be used both for supervised and unsupervised evaluation. We design a test video set specifically for evaluation of the quality of video segmentation and evaluate the proposed method using both this set and segmentations of real life videos. We compare our method against a state of the art supervised evaluation method both in supervised and unsupervised modes. The remainder of the paper is organized as follows. In Section 2, we discuss proposed criteria and metrics. In Section 3, we give a detailed overview of the proposed method. Section 4 provides evaluation and results. Finally, conclusion is discussed in Section 5.

2 PROPOSED CRITERIA AND METRICS

An initial step for any evaluation process is determining the criteria of evaluation. As discussed in the previous section, a number of researchers proposed several criteria for evaluation of image segmentation. However, there are some differences between image segmentation and video segmentation, and thus it is necessary to propose new criteria by considering additional characteristic relating to good quality of video segmentation.

For evaluating video segmentation quality, the stability of the boundaries and consistent region identity between consequent frames should be evaluated (Grundmann et al., 2010). Given this, and in the light of the criteria proposed by Haralick and Shapiro, we propose the following set of criteria:

1. The regions must be uniform, homogeneous, simple and without holes.
2. Adjacent regions should have significant differences with respect to the characteristic on which they are uniform.
3. Corresponding regions between consequent frames should be consistent.
4. Boundaries of segmented frame should be smooth, stable and accurate when compared with the bound-

aries of original frame.

This work is based on low level image features in line with previous unsupervised methods (Zhang et al., 2008). For this reason, we will not use the second criterion when evaluating the quality of video segmentation, as it is difficult to find meaningful adjacent segments without semantic information. In the next section, we will consider the metrics for measuring the quality of the video segmentation according to the remaining three criteria: those for measuring intra-region uniformity and homogeneity (criterion 1), those for measuring region consistency between consequent frames (criterion 3), and those for measuring boundary accuracy (criterion 4).

2.1 Intra-Region Uniformity

The uniformity of regions can be divided into two categories, color uniformity and texture uniformity. The former means that the pixel colors of a region should have similar values, the latter means that each region should have consistent texture. In (Zhang et al., 2008) the intra region uniformity metrics are classified into four classes: based on color error, squared color error, texture and entropy. We select two simple and easy to understand metrics, the first is F_{RC} (Rosenberger and Chehdi, 2000), which is based on squared color error and measures the intra region color disparity, and the second is texture variance *Tex_var* (Correia and Pereira, 2003), which measures the texture uniformity and is based on the variance of Y, U and V layers.

2.2 Temporal Region Consistency

There is a number of methods which can be used to evaluate consistency between two regions. In addition, a number of metrics have been designed specifically to measure the similarity between two ground truth images, such as Variation of Information (VI) (Meilă, 2003; Unnikrishnan et al., 2005), Global Consistency Error GCE (Martin et al., 2001) and probabilistic rand index PRI (Unnikrishnan et al., 2005). GCE and VI have been designed to compare two segmentations, while PRI has been designed to compare more than two segmentations. VI is an information based metric, which considers mutual information between two segmentations, whilst GCE is a region based metric, and designed to quantify the consistency between image segmentations of different granularities (Unnikrishnan et al., 2007). In video segmentation corresponding regions from consecutive frames should have consistent color and granularity. Here, we propose to evaluate consistency between two consecutive frames by combining GCE with positive correlation. Although GCE has been used for image segmentation evaluation, no one has combined it with positive correlation. Our approach to combine GCE with positive correlation has the advantage of taking into account both the consistency of region granularity and the color consistency of regions between consecutive frames.

2.3 Boundary Stability and Accuracy

To measure the accuracy between the boundaries of segmented and original frames, we will use F-measure metric, (Martin et al., 2001) which is the most popular boundary based metric for evaluation of image segmentation (Galasso et al., 2013). We develop a method to detect boundaries of original frames in case of unsupervised evaluation, and use ground truth boundaries in case of supervised evaluation. In the next Section we will explain the proposed method.

3 PROPOSED METHOD

In this work, we propose a method on the basis of the new criteria proposed in Section 2. Our method can be used both for supervised and unsupervised evaluation. The former evaluation uses the boundaries of the ground truth, the latter uses the boundaries found in the original frame, for which we use a combination of low-pass filtering to remove noise and multiscale edge detection. Our method uses the found boundaries at two different stages. First, our method uses them as a map controlling the regions of the segmented frame for measuring intra region uniformity as outlined in Section 3.1. Second, it uses the detected boundaries to evaluate the accuracy of the boundaries of the segmented frames.

3.1 Intra-Region Uniformity

Selecting semantic regions composing an image needs either human help or ground truth template, which are not available for unsupervised segmentation. We overcome this issue by detecting and using the boundaries of the original video frames. Thus our process of evaluation of intra-region uniformity consists of the following three steps.

1. *Detecting boundaries*. We use our own method relying on a combination of low-pass filtering to remove noise and multiscale edge detection.
2. *Selecting regions from segmented frame*. The detected boundaries produced from previous step

are used to select the regions of segmented frame. We will use quad-tree image decomposition to separate the segmented frame into a number of rectangular areas not containing any boundaries from the original video frame. The uniformity of each of the rectangular areas is then evaluated in the next step.

3. $Evaluating\ intra-region\ uniformity$. The selected regions produced from the previous step are evaluated using two metrics selected in Section 2.1, namely F_{RC} and Tex_var.

We will explain the metrics F_{RC} and Tex_var individually. Let N be the total number of regions of segmented image I, with the height I_x and width I_y, j is the index of regions $j \in (1,2,3,...,N)$, R_j represent set of pixels in the region j where $R_j \subset \left(\cup_{j=1}^{N}(R_j)\right)$, S_j denotes the area of the region j, $C_x(P)$ is the color intensity value for pixel P ($x \in$ red, green, or blue component), and the area of full image is $S_I = I_x \times I_y$.

The mean value of component x in region j can be defined as the following:

$$\widehat{C_x}(R_j) = \frac{1}{S_I} \sum_{P \in R_j} C_x(P) \tag{1}$$

F_{RC} is based on the squared color error and measures the intra region color disparity, squared color error can be defined as the following:

$$e_x^2(R_j) = \sum_{P \in R_j} (C_x(P) - \widehat{C_x}(R_j))^2 \tag{2}$$

F_{RC} metric can be explained as the following:

$$D(I) = \frac{1}{N} \sum_{j=1}^{N} \frac{S_j}{S_I} \times e_x^2(R_j) \tag{3}$$

where $D(I)$ is the F_{RC} color disparity of image/frame I, and $e_x^2(R_j)$ is squared color error of region R_j.
Tex_var (Correia and Pereira, 2003) is defined as the following:

$$Tex\_var(I) = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{5} \left(3 \times \sigma^2_{y(R_j)} + \sigma^2_{u(R_j)} + \sigma^2_{v(R_j)}\right) \tag{4}$$

where $Tex\_var(R_j)$ is the texture variance of the region(R_j), σ_y, σ_u and σ_v are the variance of the Y, U and V components in R_j region, respectively. Both $D(I)$ and $Tex\_var(I)$ metrics are normalized to intra region uniformity I_U and texture uniformity T_U respectively, by the following function:

$$\eta = \left(\frac{1}{1 - \frac{\nu^{0.5}}{128}} - 0.5\right) \times 2 \tag{5}$$

where η and ν represent normalized value (between 0 and 1) and initial value (between 0 and 128^2) of the metrics respectively.

Figure 1: Result of boundary detection for frame number one, from left to right: Soccer sequence and Ice sequence.

A real scene can consist of both color and texture regions, and it is difficult to decide which region category is popular in the scene, it is totally application based. For this reason, both color uniformity and texture uniformity to measure the region uniformity are selected, and we take their average to take them both into account.

3.2 Region Consistency

The content of consecutive frames in video sequences is usually not exactly the same, but still there is consistency and similarity between them, which is dependent on the complexity of the sequence. As we discussed previously, according to the listed criteria in Section 2, identical regions between consequent frames should be consistent. Global consistency error GCE (Unnikrishnan et al., 2007) was used to compute the global consistency error between approximately similar images. We will use global consistency error GCE to compute the global consistency index GCI. In this work, for evaluating region consistency, we want to ensure the consistency according to both metrics. To obtain this, we will select the minimum value between GCI and positive correlation between consequent frames of segmented video. GCI can be explained as follows.

Let $\setminus$ denote set difference, and $|x|$ the cardinality of set x. Let S_1 and S_2 be two segmentations. For a given pixel p_i, consider the segments that contain p_i in S_1 and S_2. We denote these sets of pixels by $R(S_1, p_i)$ and $R(S_2, p_i)$ respectively, the local refinement error is defined as:

$$E(S_1, S_2, p_i) = \frac{|R(S_1, p_i) \setminus R(S_2, p_i)|}{|R(S_1, p_i)|} \tag{6}$$

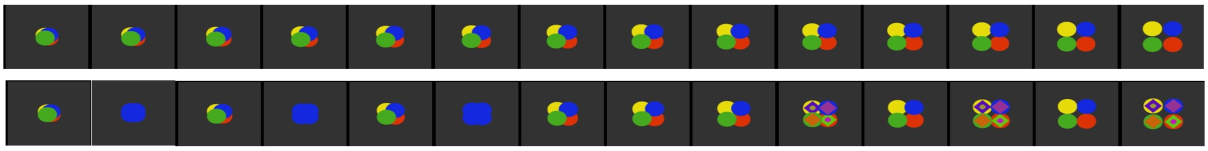
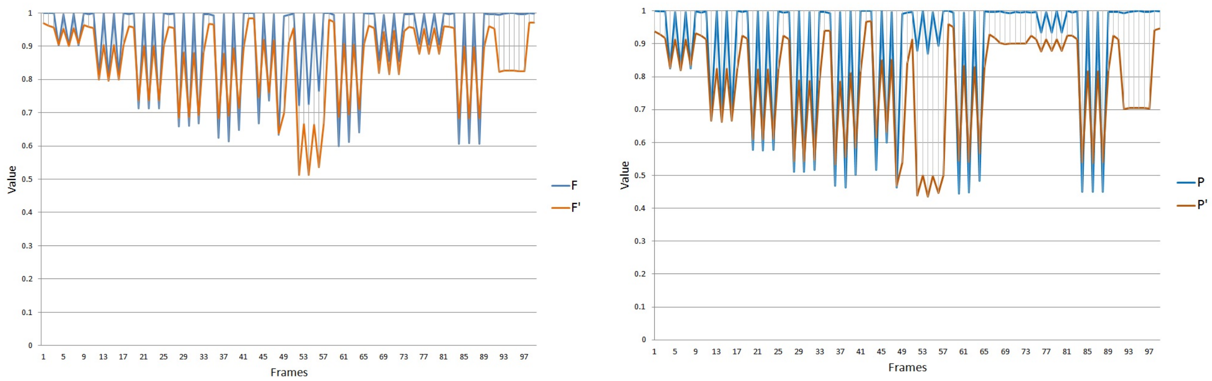


Figure 2: Sample of the synthetic video, first row is the ground truth and the second is segmented, both sequences are in the same order.



(a) (b)

Figure 3: Result of: (a) F and F′; (b) P and P′.

$$GCE(S_1,S_2) = \frac{1}{n} min\left(\sum_i E(S_1,S_2,p_i), \sum_i E(S_2,S_1,p_i)\right) \quad (7)$$

$GCE(S_1,S_2)$ is the global consistency error between frame S_1 and S_2, and n is the number of pixels.

$$GCI(S_1,S_2) = 1-(GCE(S_1,S_2)) \quad (8)$$

In our case, we applied the GCI to each frame, which means that S_1 and S_2 represent two consecutive frames.

The positive correlation between consecutive frames can be calculated as the following:

$$Corr(S_1,S_2) = \begin{cases} r(S_1,S_2) & r(S_1,S_2) \geq 0 \\ 0 & r(S_1,S_2) < 0 \end{cases} \quad (9)$$

where $r(S_1,S_2)$ is the Pearson's correlation between frame S_1 and S_2, and can be defined as the following:

$$r(S_1,S_2) = \frac{n\sum(S_1S_2)-(\sum S_1)(\sum S_2)}{\sqrt{(n\sum S_1^2-(\sum S_1)^2)(n\sum S_2^2-(\sum S_2)^2)}} \quad (10)$$

3.3 Boundary Assessment

F-measure (Martin et al., 2001) is the most popular metric in this area, as we discussed in Section 2, we will use F to refer to F-measure.

$$F = \frac{2*P*R}{P+R} \quad (11)$$

where P is the precision of the boundaries, and R is the recall of the boundaries.

3.4 Combining Metrics

The selected metrics explained in the previous sections are combined as a version of F formula. F is the harmonic mean of precision and recall, with precision penalising oversegmentation and recall penalising undersegmentation, both of which are important for evaluating the quality of video segmentation. In this work we will update precision P to P′ to include the metrics evaluating region uniformity and consistency, which also play important role in penalising over- and under-segmentation.

$$F' = \frac{2*P'*R}{P'+R} \quad (12)$$

$$P' = \frac{P+\alpha}{2} \quad (13)$$

where P' is the updated precision, and it is the average between precision P and α. We define α as:

$$\alpha = \frac{2*U*C}{U+C} \quad (14)$$

where α is the harmonic mean between intra-region uniformity U and consistency C. We will define them as the following:

$$U = \frac{I_U+T_U}{2} \quad (15)$$

$$C = min(GCI, Corr) \quad (16)$$

where I_U is the minimum value of normalized intra region uniformity among R, G, and B layers, obtained

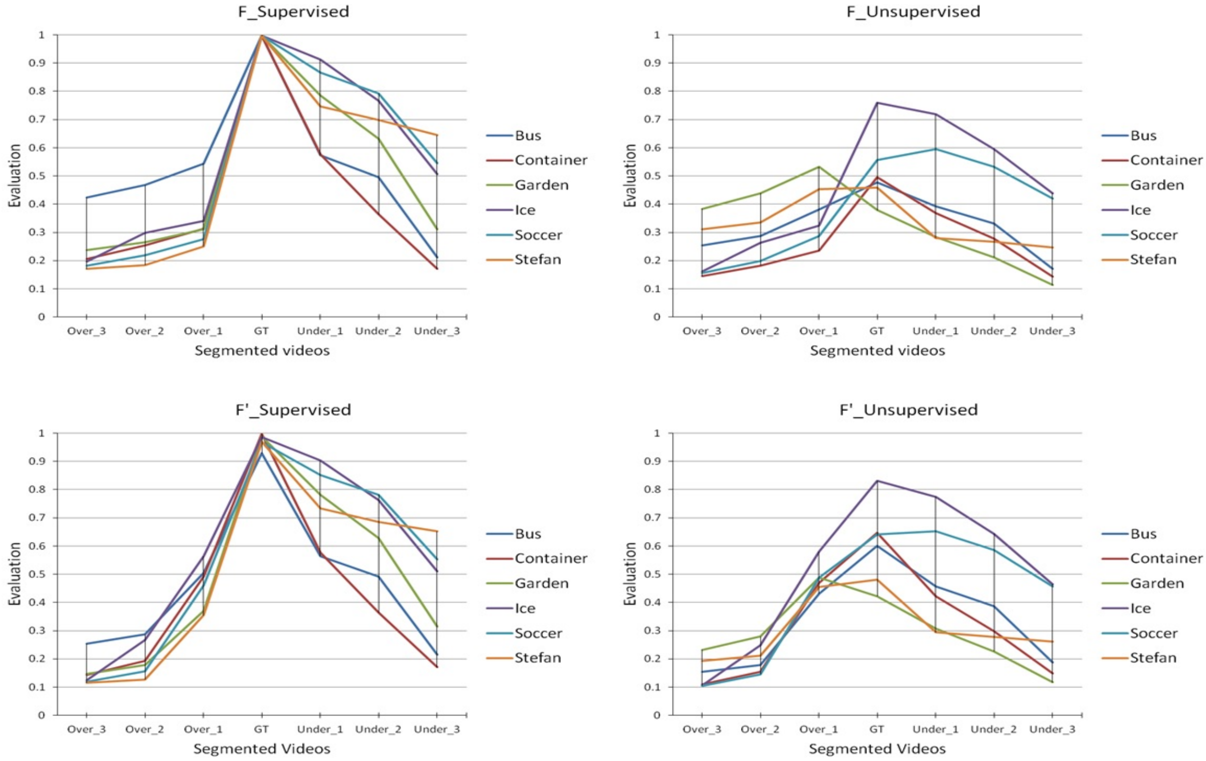


Figure 4: Segmentation evaluation result using F and F′ for the supervised and unsupervised cases, for six video sequences (Bus, Container, Garden, Ice, Soccer and Steven). Each video has seven degrees of segmentation, ground truth, three degrees of under segmentation and three degrees of oversegmentation.

using Eqs. 3 and 5, T_U is the normalized texture uniformity obtained using Eqs.4 and 5, *GCI* is the minimum value of global consistency index among R, G, and B layers (Eq. 8), and *Corr* is the minimum value of positive correlation among R, G, and B layers (Eq. 9).

4 EVALUATION AND RESULTS

4.1 Synthetic Data

We created two synthetic videos of length 100 frames each. The first video depicts for differently colored circles moving from different corners towards each other, meeting in the middle and then moving to the opposite corners (Figure 2). The second video represents different defects in the segmentation of the first video, such as over- and undersegmentation, undetected objects, inconsistent object identity (swapping of identity between objects), etc. We also insert "correctly segmented" frames between the "defective segmentations" to represent inconsistent temporal segmentations. The dataset is available by contacting the authors.

4.2 Real Video Data

The real video dataset is from (Chen and Corso, 2010) and is a subset of the *Xiph.org* videos. The selected dataset used in this work can be divided into three groups: ground truth, oversegmented and undersegmented. We selected six different videos labelled with a 24-class semantic pixel labeling as a ground truth (Chen and Corso, 2010). For each video, from ground truth frames, we created three degrees of undersegmentation. Also for each video, we created three degrees of oversegmentation using the hierarchical graph-based method (Grundmann et al., 2010). The length of the videos varied from 69 to 86 frames.

4.3 Results on Synthetic Video

This example explains the ability of F and F′ to evaluate different types of segmentation defects. We applied F and F′ to the synthetic video data set described in the previous section. The results of F are accurate in most of the cases, but it is not as strict as the F′

in penalising inconsistent object identity and undersegmention. Figure 3 shows the differences between F and F′, and P and P′. The frames between number 53 to 57 are undersegmented and frames between 94 and 97 contain inconsistent object identity.

4.4 Results on Real Video

Alongside the synthetic video we evaluated our method on six real videos, as outlined in Section 4.2. For each real video from the dataset we created seven segmentations of different quality, containing three degrees of oversegmentation, ground truth, and three degrees of undersegmentation (Figure 6). Figures 4 and 5 provide the comparative information on F and F′ over these segmentations. Whilst F and F′ are approximately the same for the undersegmented and ground truth, their evaluation behavior for the oversegmented areas is different, F′-measures is more consistent with the perceptual quality of the segmentations *Over*_3, *Over*_2 and *Over*_1 than F, we can observe significant difference in the quality of *Over*_2 and *Over*_1 (Figure 4 and 6).

5 CONCLUSION

The main contributions of this work are the proposal of the new criteria of good video segmentation quality, proposal of a new evaluation method based on this criteria, which can be used both for supervised and unsupervised evaluation, and creation of a synthetic video test set specifically for the purpose of evaluation of the performance of the proposed method. The results showed that our method can evaluate video sequences quality better than F, on different types of content. This is due to our method taking into account region uniformity and consistency between consecutive frames, which we included in the new criteria.

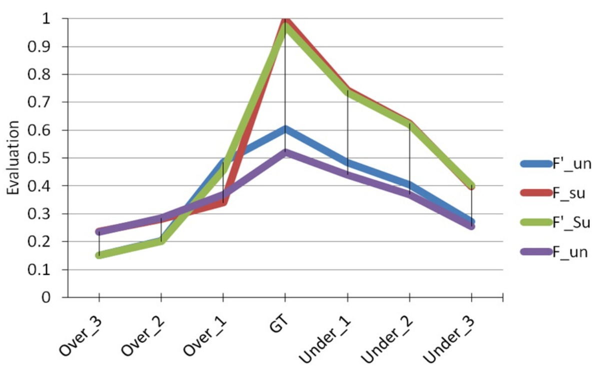


Figure 5: Result of real video evaluation using F and F′ evaluation metrics, for the supervised and unsupervised case. We took the average score of all videos for each segmentation quality.

REFERENCES

Borsotti, M., Campadelli, P., and Schettini, R. (1998). Quantitative evaluation of color image segmentation results. *Pattern Recognition Letters*, 19(8):741 – 747.

Chabrier, S., Emile, B., Rosenberger, C., and Laurent, H. (2006). Unsupervised performance evaluation of image segmentation. *EURASIP Journal on Applied Signal Processing*, 2006:217–217.

Charron, C. and Hicks, Y. (2010). An evolving mog for online image sequence segmentation. In *ICIP*, pages 2189–2192.

Chen, A. Y. C. and Corso, J. J. (2010). Propagating multiclass pixel labels throughout video frames. In *Proc. of Western New York Image Processing Workshop*.

Chen, H.-C. and Wang, S.-J. (2004). The use of visible color difference in the quantitative evaluation of color image segmentation. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 3:iii–593–6 vol.3.

Correia, P. and Pereira, F. (2003). Objective evaluation of video segmentation quality. *IEEE Trans. on Image Processing*, 12(2):186–200.

Dey, V., Zhang, Y., and Zhong, M. (2010). A review on image segmentation techniques with remote sensing perspective. In *Proc. of the Int. Society for Photogrammetry and Remote Sensing Symposium (ISPRS)*, volume 38, pages 5–7.

Erdem, Ç. E., Sankur, B., and Tekalp, A. M. (2004). Performance measures for video object segmentation and tracking. *IEEE Trans. on Image Processing*, 13(7):937–951.

Galasso, F., Nagaraja, N. S., Cárdenas, T. J., Brox, T., and Schiele, B. (2013). A unified video segmentation benchmark: Annotation, metrics and analysis. In *IEEE Int. Conf. on Computer Vision (ICCV)*.

Gelasca, E. D. and Ebrahimi, T. (2006). On evaluating metrics for video segmentation algorithms. In *Second Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*.

Greenspan, H., Dvir, G., and Rubner, Y. (2004). Context-dependent segmentation and matching in image databases. *Computer Vision and Image Understanding*, 93(1):86–109.

Grundmann, M., Kwatra, V., Han, M., and Essa, I. (2010). Efficient hierarchical graph-based video segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2141–2148.

Haralick, R. M. and Shapiro, L. G. (1985). Image segmentation techniques. *Computer vision, graphics, and image processing*, 29(1):100–132.

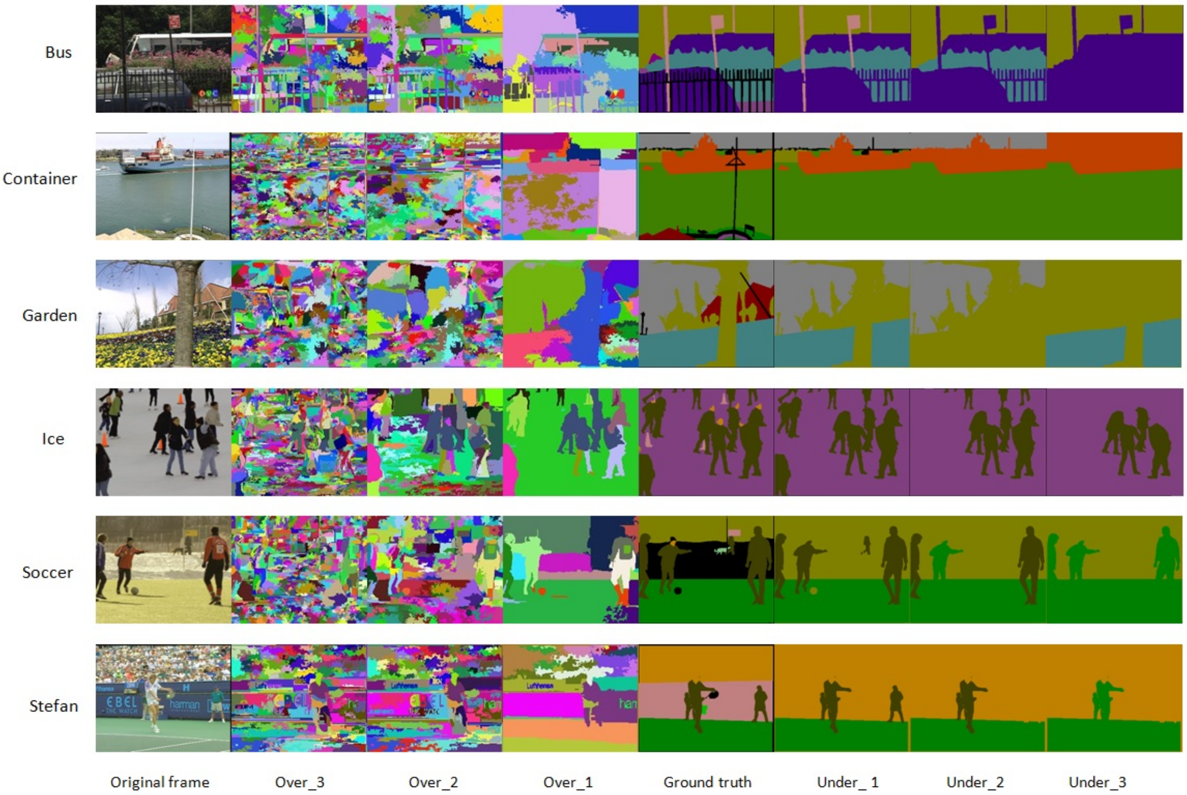


Figure 6: Visual comparative of the seven different segmentation quality.

Kaloskampis, I. and Hicks, Y. (2014). Estimating adaptive coefficients of evolving gmms for online video segmentation. In *6th Int. Symp. on Communications, Control and Signal Processing (ISCCSP)*, pages 513–516.

Levine, M. D. and Nazif, A. M. (1985). Dynamic measurement of computer generated image segmentations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, (2):155–164.

Liu, J. and Yang, Y.-H. (1994). Multiresolution color image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(7):689–700.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. Eighth IEEE Int. Conf. on Computer Vision (ICCV)*, volume 2, pages 416–423.

Meilă, M. (2003). Comparing clusterings by the variation of information. In *Learning theory and kernel machines*, pages 173–187.

Morris, O., Lee, M. d. J., and Constantinides, A. (1986). Graph theory for image analysis: an approach based on the shortest spanning tree. *Communications, Radar and Signal Processing, IEE Proc. F*, 133(2):146–152.

Rosenberger, C. and Chehdi, K. (2000). Genetic fusion: application to multi-components image segmentation. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 6, pages 2223–2226.

Unnikrishnan, R., Pantofaru, C., and Hebert, M. (2005). A measure for objective evaluation of image segmentation algorithms. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition-Workshops (CVPR) Workshops*, pages 34–34.

Unnikrishnan, R., Pantofaru, C., and Hebert, M. (2007). Toward objective evaluation of image segmentation algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(6):929–944.

Xu, C. and Corso, J. J. (2012). Evaluation of super-voxel methods for early video processing. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1202–1209.

Zhang, H., Fritts, J. E., and Goldman, S. A. (2004). An entropy-based objective evaluation method for image segmentation. In *Electronic Imaging*, pages 38–49.

Zhang, H., Fritts, J. E., and Goldman, S. A. (2008). Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, 110(2):260–280.