STANDALONE OBJECTIVE EVALUATION OF SEGMENTATION QUALITY

Paulo Correia, Fernando Pereira
Instituto Superior Técnico - Instituto de Telecomunicações
Av. Rovisco Pais, 1049-001 Lisboa, PORTUGAL
E-mail: Paulo.Correia@lx.it.pt, Fernando.Pereira@lx.it.pt

ABSTRACT

The identification of objects in video sequences, i.e. video segmentation, plays a major role in emerging multimedia interactive services, such as those enabled by the ISO MPEG-4 and MPEG-7 standards. In this context, assessing the adequacy of the identified objects to the application targets, i.e. the evaluation of segmentation quality, assumes a crucial importance.

Video segmentation technology has received considerable attention in the literature, and several algorithms have been proposed to address various types of applications. However, the segmentation quality performance evaluation of those algorithms is often adhoc, and a well-established solution is not available. In fact, the field of objective segmentation quality evaluation is still maturing, and recently some efforts have been made, mainly following the MPEG object-based coding and description developments.

This paper discusses the problem of objective segmentation quality evaluation in its most difficult scenario: standalone evaluation, i.e. when a reference segmentation is not available for comparative evaluation. In particular, objective metrics are proposed for the evaluation of standalone segmentation quality for both individual objects and the overall segmentation partition.

1. INTRODUCTION

With the recent publication of the MPEG-4 standard [3], allowing to independently encode audiovisual objects, and the development of the MPEG-7 standard [7], allowing the content-based description of audiovisual material, the MPEG committee has given a significant contribution for the development of a new generation of interactive multimedia services. Innovative types of interaction are often based on the understanding of a video scene as composed by a set of video objects, to which it is possible to associate specific information as well as interactive “hooks” to deploy the desired application behaviour.

To enable such type of interactive services, an understanding of the scene semantics is required, notably in terms of the relevant objects that are present. It is in this context that video segmentation plays a determinant role. Segmentation may be automatically obtained at the video production stage, e.g. when using chroma keying techniques, or it may have to be directly obtained from the images captured by a camera through the usage of appropriate segmentation algorithms.

The evaluation of the adequacy of a segmentation algorithm, and its parameters’ configuration, for a given application can be crucial to guarantee that the application interactive requirements can be fulfilled.

The current practice for segmentation quality evaluation mainly consists in the subjective adhoc assessment, by a representative group of human viewers. This is a time-consuming and expensive process, whose subjectivity can be minimised by following strict evaluation conditions, with the video quality evaluation recommendations developed by ITU providing valuable guidelines [4, 5].

Alternatively, objective segmentation quality evaluation methodologies can be used, even if the amount of attention devoted to this issue is not comparable to the investment on the segmentation algorithms themselves. Some proposals for objective evaluation have been made since the 1970’s, mainly for assessing the performance of edge detectors [11]. More recently, the emergence of the MPEG-4 and MPEG-7 standards, has given a new impulse not only to the segmentation technology, but also to the segmentation quality evaluation methodologies – see for instance [8, 10]. However, the available metrics for segmentation quality evaluation typically perform well only for very constrained applications scenarios.

This paper discusses the objective evaluation of segmentation quality, in particular when no ground truth segmentation is available to use as a reference for comparison: standalone evaluation.

The various types of standalone segmentation quality evaluation are discussed in Section 2. Metrics for individual object and overall segmentation quality evaluation are proposed in Sections 3 and 4, respectively. Results are presented in Section 5 and conclusions in Section 6.

2. STANDALONE SEGMENTATION EVALUATION

Standalone segmentation quality evaluation is performed when no reference segmentation is available. Therefore, the a priori information that may be available about the expected segmentation results has a decisive impact on the type of evaluation to be performed, so that meaningful results can be
achieved. In particular, standalone evaluation of segmentation quality is not expected to provide as reliable results as the evaluation relative to a reference segmentation. A discussion on the relative evaluation of segmentation quality evaluation can be found in [2].

When performing segmentation quality evaluation, two types of measurements can be targeted:

- **Individual object evaluation** – When one of the objects identified by the segmentation algorithm is independently evaluated in terms of its segmentation quality.

- **Overall evaluation** – When the complete set of objects identified by the segmentation algorithm are globally evaluated as the components of the video sequence partition.

Objective segmentation quality evaluation uses automatic tools and thus produces objective evaluation measures. The automatic tools operate on segmentation results obtained for a selected set of sequences and, in the case of individual object evaluation, the object whose segmentation quality is to be assessed has first to be selected.

Overall segmentation quality evaluation requires the estimation of individual object quality, and the weighting of those values according to each object’s relevancy in the scene, since segmentation errors in the more important objects are more noticeable to a human viewer. Additionally, the correct detection of the target objects should be checked.

Individual object segmentation quality evaluation is valuable when objects are independently manipulated, e.g. for reusing in different contexts. On the other hand, the overall segmentation quality evaluation may determine whether the segmentation algorithm is adequate for the application addressed.

Both the individual object and the overall segmentation quality measures can be computed for each time instant, requiring that some temporal processing of the instantaneous results is after done to reflect the segmentation quality over the complete sequence or shot. For instance, a temporal mean or median may be computed.

Building on the existing knowledge on segmentation quality evaluation and also on some relevant aspects from the video quality evaluation field, a set of relevant features to be assessed has first to be selected.

With individual object segmentation quality evaluation, significant assessment results are only expected for well-constrained applications, and these results mainly provide qualitative information for the ranking of segmentation partitions and algorithms.

### 3. INDIVIDUAL OBJECT EVALUATION

Metrics for individual object standalone segmentation quality evaluation can be established based on the expected homogeneity of each object’s features (**intra-object features**), as well as on the observed differences of some key features against those of the neighbours (**inter-object features**).

#### Intra-object

Homogeneity can be evaluated by means of spatial and temporal object features, as discussed below.

The **spatial features** considered for individual object evaluation, and corresponding metrics, are:

- **Shape regularity** – Regularity of shapes can be evaluated by geometrical features such as the compactness (**compact**), or a combination of circularity and elongation (**circ_elong**) of the objects:

  $$
  \text{compact}(E) = \max \left( \frac{\text{perimeter}^2(E)}{75 \cdot \text{area}(E)}, 1 \right)
  $$

  $$
  \text{circ}_{\text{elong}}(E) = \max \left( \text{circ}(E), \max \left( \frac{\text{elong}(E)}{5}, 1 \right) \right)
  $$

  With circularity and elongation defined by:

  $$
  \text{circ}(E) = 4 \pi \cdot \frac{\text{area}(E)}{\text{perimeter}^2(E)},
  \text{elong}(E) = \frac{\text{area}(E)}{2 \cdot \text{thickness}(E)}
  $$

  Here \(\text{thickness}(E)\) is the number of morphological erosion steps that can be applied to the object until it disappears. The normalizing constants were empirically determined after an exhaustive set of tests.

- **Spatial uniformity** – Spatial uniformity can be evaluated by features such as spatial perceptual information (SI) [5], and texture variance (text_var) – see for instance [6].

#### Temporal features

The **temporal features**, and corresponding metrics, considered are:

- **Temporal stability** – A smooth temporal evolution of object features can be tested for checking temporal stability. These features include: size, position, temporal perceptual information [5], criticality [9], texture variance, circularity, elongation and compactness. The selected metrics for temporal stability evaluation are:

  $$
  \text{size}_{\text{diff}} = |\text{area}(E_i) - \text{area}(E_{i-1})|
  $$

  $$
  \text{elong}_{\text{diff}} = |\text{elong}(E_i) - \text{elong}(E_{i-1})|
  $$

  $$
  \text{crit}_{\text{diff}} = |\text{crit}(E_i) - \text{crit}(E_{i-1})|
  $$

  With \(\text{crit}(E)\) being the criticality value as defined in [9].

- **Motion uniformity** – The uniformity of motion can be evaluated by features such as the variance of the object's motion vector values (mot_var), or by criticality (crit).

  The above spatial and temporal features are not expected to be homogeneous for every segmented object; the applicability and importance of the corresponding metrics is conditioned by the type of application addressed.

#### Inter-object features

Give an indication if the objects were correctly identified as separate entities. These features can be computed either locally along the object boundaries, or for the complete object area. Again these features may be applicable only in some circumstances, such as when a significant contrast, or some feature value difference, between neighbouring objects is expected.
• **Local contrast to neighbours** – A local contrast metric can be used for evaluating if a significant contrast between the inside and outside of an object, along the object border, exists:

\[
\text{contrast} = \frac{1}{4 \times 255 \times N_i} \sum_{j \in \mathcal{N}_i} \left(2 \times \max(DY_{ij}) + \max(DU_{ij}) + \max(DV_{ij})\right)
\]

Where \(N_i\) is the number of border pixels for the object and \(DY_{ij}\), \(DU_{ij}\), and \(DV_{ij}\) are the differences between an object’s border pixel \(Y\), \(U\) and \(V\) components, respectively, and its neighbours.

- **Differences between neighbouring objects** – Several features, for which objects are expected to differ from their neighbours, can be tested. Examples are the shape regularity, spatial uniformity, temporal stability, and motion uniformity values, whenever each of them is relevant taking the application characteristics into account. In particular a metric for the motion uniformity feature is considered of interest:

\[
\text{mot\_unif}_{i,syb\_adj} = \frac{1}{N} \sum_{j \in \mathcal{N}_i} \text{mot\_unif}_j - \text{mot\_unif}_i
\]

Where \(i\) is the object under analysis, \(N\) and \(\mathcal{N}_i\) are, respectively, the number and the set of neighbours of object \(i\), and the motion uniformity for each object is computed as:

\[
\text{mot\_unif}_i = \text{mot\_var} + \text{crit}
\]

Each of the elementary metrics considered for individual object segmentation quality evaluation is normalized to produce results in the interval \([0, 1]\), with the highest values associated to the best segmentation quality results.

Since the usefulness of the various standalone evaluation elementary metrics has a strong dependency on the characteristics of the type of content/application considered, a single general-purpose composite metric cannot be established. Instead, the approach taken here is to select two major classes of content differing in terms of their spatial and temporal characteristics, and propose different composite metrics for each of them.

The two classes of content selected are:

- **Content class I: stable content** – Relevant for applications which content is temporally stable and have reasonably regular shapes. Additionally, the contrast between objects is expected to be strong.

- **Content class II: moving content** – Relevant for applications which content motion is rather important. Consequently, temporal stability is less relevant, motion uniformity is more significant and neighbouring objects may be spatially less contrasted, while their motion differences are more noteworthy. Regular shapes are still expected, even if assuming a lower importance.

### 3.1. Individual Object Metric for Stable Content

For content class I, stable content, a composite metric is proposed that excludes the elementary metrics related to spatial uniformity, as arbitrary spatial patterns may be found in the expected objects, and to motion uniformity, as motion is not very relevant in this case. Thus, the classes of elementary metrics considered for standalone individual evaluation of stable content are:

- **Shape regularity** – Two elementary metrics, compactness (compact) and a combination of circularity and elongation (circ_elong), are considered for evaluation of the shape regularity class.

- **Temporal stability** – Elementary metrics for the stability of size (size_adj), elongation (elong_adj) and criticality (crit_adj) are used to evaluate this class of metrics.

- **Local contrast to neighbours** – A local contrast metric (contrast) is considered for the evaluation of the contrast between neighbouring objects.

The proposed composite metric for standalone evaluation of segmentation quality, for content class I, \(\text{Seg\_qual\_std\_stable}\) is the temporal average of the corresponding instantaneous values of \(\text{Inst\_seg\_qual\_std\_stable}\), given by:

\[
\text{Inst\_seg\_qual\_std\_stable} = \text{intra} + \text{inter}
\]

With:

\[
\text{intra} = 0.30 \times (0.5 \times \text{circ\_elong} + 0.5 \times \text{compact}) + 0.33 \times (0.33 \times \text{size\_adj} + 0.33 \times \text{elong\_adj} + 0.33 \times \text{crit\_adj})
\]

and:

\[
\text{inter} = 0.37 \times \text{contrast}
\]

### 3.2. Individual Object Metric for Moving Content

For content class II, the composite metric again includes only the relevant classes of elementary metrics. In this case, the content is not expected to be temporally stable, but the objects should have uniform motion, and the neighbouring objects motion differences should be pronounced. The classes of metrics considered for the standalone segmentation quality evaluation of this type of content are:

- **Shape regularity** – The same elementary metrics, compact and circ_elong, are again used, even if, due to motion, the shape regularity assumption may sometimes not be completely verified.

- **Motion uniformity** – The criticality metric (crit) is used to evaluate whether objects exhibit a reasonably uniform motion.

- **Local contrast to neighbours** – Even if contrast is not so important in terms of segmentation quality evaluation as for stable content, the local contrast (contrast) metric is yet considered useful.

- **Difference between neighbouring objects** – Since neighbouring objects are expected to exhibit different motion characteristics, the motion uniformity difference metric (mot_unif_adj) is used.

The proposed composite metric for content class II, \(\text{Seg\_qual\_std\_moving}\) is the temporal average of the corresponding instantaneous values \(\text{Inst\_seg\_qual\_std\_moving}\), given by:

\[
\text{Inst\_seg\_qual\_std\_moving} = \text{intra} + \text{inter}
\]
4. OVERALL SEGMENTATION EVALUATION

The objective overall segmentation quality evaluation combines the individual evaluation of each object’s segmentation quality, with the corresponding relevance value and a factor reflecting the similarity between the target and the estimated objects.

Individual object evaluation has been specified in the previous Section. The relevance of objects is evaluated using a metric called Relevance\_context, which has been proposed in [1]. This metric computes a relevance value reflecting how much the human viewer attention is attracted by a given object, and produces results in the [0,1] range, with the restriction that the relevancies of all objects composing a partition at a given time instant sum to one. Value one corresponds to the highest possible relevance.

The assessment of the similarity of objects for standalone segmentation quality evaluation, and the computation of the overall segmentation quality metric are described below.

4.1. Similarity of Objects Evaluation

The similarity of objects is evaluated by computing a metric called Sim\_obj\_factor, which is a multiplicative factor to be included in the overall segmentation quality evaluation metric.

In standalone evaluation, this similarity is mainly reduced to an evaluation about the correctness of the number of objects detected, if this information is available. The corresponding metric (num\_obj\_comparison) is defined by:

\[
\text{num\_obj\_comparison} = \frac{\min(\text{num\_est\_obj}, \text{num\_target\_obj})}{\max(\text{num\_est\_obj}, \text{num\_target\_obj})}
\]

Where num\_est\_obj and num\_target\_obj are the numbers of estimated and target objects, respectively.

This metric provides a limited amount of information, in particular not distinguishing between too many or too few detected objects. To make the Sim\_obj\_factor metric more informed, it is possible to consider also a measure of the number of objects stability (num\_obj\_stability), applicable whenever the evolution of the segmentation partition is assumed to be smooth:

\[
\text{num\_obj\_stability} = \frac{\min(\text{num\_obj}_t, \text{num\_obj})}{\max(\text{num\_obj}_t, \text{num\_obj})}
\]

Where num\_obj\_t is the number of estimated objects in time instant \( t \).

The proposed Sim\_obj\_factor metric for standalone segmentation quality evaluation is thus obtained by complementing the num\_obj\_comparison factor with the num\_obj\_stability factor:

\[
\text{Sim\_obj\_factor} = \text{num\_obj\_comparison} \cdot \text{num\_obj\_stability}
\]

Whenever one of the two factors above cannot be computed, or is not applicable, only the other is considered. Additionally, since the two factors vary as time evolves, a Sim\_obj\_factor representative of the complete sequence is obtained by a temporal average of the instantaneous values.

4.2. Overall Segmentation Quality Metric

The computation of the overall segmentation quality metric, both for standalone and relative evaluation, combines the appropriate measures of individual object quality, their relevance and the similarity of objects factor. The proposed metric is computed by:

\[
\text{Seg\_qual} = \text{Sim\_obj\_factor} \cdot \left( \sum \left( \text{Seg\_qual\_ind}(E_i) \cdot \text{Relevance\_context}(E_i) \right) \right)
\]

Where Seg\_qual\_ind\( (E_i) \) is the individual segmentation quality for object \( i \), Relevance\_context\( (E_i) \) is the corresponding relevance, and Sim\_obj\_factor is the factor evaluating the correspondence between the detected and target objects. The sum is performed for all the estimated objects.

Instead of including the temporal dimension influence separately in each factor, as presented so far, a weighting of the instantaneous objects’ quality by their instantaneous relevance values, and then by the instantaneous similarity of objects factor, to reflect the variations in quality, relevance or similarity values that may occur along time, is used.

With this overall segmentation quality evaluation metric, the higher the individual object quality is for the most relevant objects, the better is the resulting quality evaluation. Therefore, the most relevant objects, which are the most visible to the human observers, have a larger impact on the overall segmentation quality evaluation. Furthermore, if a correct match between target and estimated objects is not achieved, then a penalizing factor is correspondingly included.

5. RESULTS

Results obtained with the metrics proposed in the previous Sections for standalone segmentation quality evaluation are discussed below, after presenting a set of test sequences and corresponding segmentation partitions.

5.1. Test Sequences and Segmentation Partitions

A series of tests of the proposed segmentation quality evaluation metrics has been performed, using several test sequences, mainly from the MPEG-4 test set, showing different spatial complexity and temporal activity characteristics. For each sequence, several segmentation partitions with different segmentation qualities were considered.

Two subsets of the test sequences, each with 30 representative images of the desired object behaviour and characteristics, are used to illustrate the obtained results. These subsequences are:
• **Akiyo**, images 0 to 29 – This is a sequence with low temporal activity and not very complex texture. It contains two objects of interest: the woman, and the background.

• **Stefan**, images 30 to 59 – This is a sequence with high temporal activity and relatively complex texture. It contains two objects of interest: the tennis player, and the background.

Sample original images and segmentation partitions are shown in Figure 1 and in Figure 2, respectively for the Akiyo and Stefan sequences. The segmentation partitions labelled as *reference* are those made available by the MPEG group, and the other partitions were created with different segmentation quality levels, ranging from a close match with the reference to more objectionable segmentations.

5.2. Standalone Segmentation Quality Evaluation Results

Standalone segmentation quality evaluation metrics are applicable only in certain circumstances, and thus the two metrics proposed have been tested with the appropriate contents. Results for these metrics, considering both the individual object and the overall evaluation cases are included below.

A set of preliminary experiments showed that similar segmentation quality evaluation results are produced independently of the input format, and thus the QCIF resolution was used to limit the algorithm execution time.

For each test sequence, the results include a graph, representing the temporal evolution of the overall segmentation quality, and a table, containing the temporal average of the instantaneous results computed both for individual object and for overall segmentation quality evaluation.

Content class I corresponds to video sequences which have relatively simple shapes, and present a limited amount of motion. To evaluate this type of content, the Akiyo test sequence and the corresponding segmentation partitions were used.

The results, included in Figure 3, show that for the woman object there are three segmentation quality groups: best quality is achieved by the reference, segmentation 1, and segmentation 2, then segmentation 3 achieves intermediate values, and finally segmentation 4 gets the worst results. These results agree generally agree with those a human observer would produce. The reference segmentation does not get the best result since a part of the woman’s hair is intensely illuminated, and when included as part of the woman it leads to a lower contrast to the background than when it is omitted, as it happens with segmentations 1 and 2. Segmentation 4, for which the woman object captures a significant part of the background, is clearly identified as the worst segmentation.

<table>
<thead>
<tr>
<th>Average Segment Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back.</td>
</tr>
<tr>
<td>Ref</td>
</tr>
<tr>
<td>Seg1</td>
</tr>
<tr>
<td>Seg2</td>
</tr>
<tr>
<td>Seg3</td>
</tr>
<tr>
<td>Seg4</td>
</tr>
</tbody>
</table>

Figure 3 - Standalone overall and individual object quality evaluation results for the Akiyo sequence
Content class II corresponds to more complex video content than in the previous case. Object shapes may not be so simple, and motion may be more important. The Stefan sequence and the corresponding segmentation partitions were used to evaluate the proposals made in this paper with this type of content.

The results, included in Figure 4, show that segmentation 1 gets the best overall segmentation quality result followed by a group formed by the reference and segmentations 2 and 3. Segmentation 4 gets the worst result. These results can be explained as follows. Segmentation 1 is more precise than the reference partition, as the reference is smoother and sometimes includes small parts of the background as belonging to the player object. The reference and segmentation 2 are correctly classified as the next quality group, but segmentation 3 should be ranked as having lower quality than these two. Finally, segmentation 4 is correctly ranked as the worst, since it is static and includes a large amount of background into the player object, but its quality value is higher than expected.

<table>
<thead>
<tr>
<th></th>
<th>Back.</th>
<th>Player</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>0.33</td>
<td>0.43</td>
<td>0.38</td>
</tr>
<tr>
<td>Seg1</td>
<td>0.34</td>
<td>0.49</td>
<td>0.42</td>
</tr>
<tr>
<td>Seg2</td>
<td>0.32</td>
<td>0.43</td>
<td>0.38</td>
</tr>
<tr>
<td>Seg3</td>
<td>0.34</td>
<td>0.45</td>
<td>0.39</td>
</tr>
<tr>
<td>Seg4</td>
<td>0.32</td>
<td>0.37</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Figure 4 - Standalone overall and individual object quality evaluation results for the Stefan sequence

The above examples demonstrate that the standalone segmentation quality evaluation algorithm is capable of ranking the quality of the various segmentation partitions, but the results must be interpreted in a rather qualitative and relative way (e.g. for ranking purposes). Standalone evaluation results are not expected to be as reliable as those of relative evaluation, but they can still be useful for identifying several quality groups among the various tested segmentations/algorithms.

6. CONCLUSIONS

Segmentation quality evaluation is a key element whenever the identification of a set of objects in a video sequence is required since it allows the assessment of the performance of segmentation algorithms in view of a given application targets.

Since a satisfying solution for segmentation quality evaluation is not yet available, this paper discusses the problem, in particular when a reference segmentation playing the role of ground truth is not available – standalone evaluation.

Metrics for both individual object and for overall segmentation quality evaluation were proposed. As expected, standalone evaluation revealed itself sensitive to the type of application/content considered. The various classes of elementary metrics available are not universally applicable, but when carefully selected metrics are employed for given classes of content then useful segmentation quality evaluation results can be obtained. Nevertheless, these evaluation results have a more qualitative rather than quantitative value, mainly allowing relative comparisons of segmentation results.

7. REFERENCES