Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 82195, Pages 1–11
DOI 10.1155/ASP/2006/82195

Video Object Relevance Metrics for Overall Segmentation Quality Evaluation

Paulo Correia and Fernando Pereira

Instituto Superior Técnico – Instituto de Telecomunicações, Av. Rovisco Pais, 1049-001 Lisboa, Portugal

Received 28 February 2005; Revised 31 May 2005; Accepted 31 July 2005

Video object segmentation is a task that humans perform efficiently and effectively, but which is difficult for a computer to perform. Since video segmentation plays an important role for many emerging applications, as those enabled by the MPEG-4 and MPEG-7 standards, the ability to assess the segmentation quality in view of the application targets is a relevant task for which a standard, or even a consensual, solution is not available. This paper considers the evaluation of overall segmentation partitions quality, highlighting one of its major components: the contextual relevance of the segmented objects. Video object relevance metrics are presented taking into account the behaviour of the human visual system and the visual attention mechanisms. In particular, contextual relevance evaluation takes into account the context where an object is found, exploiting for instance the contrast to neighbours or the position in the image. Most of the relevance metrics proposed in this paper can also be used in contexts other than segmentation quality evaluation, such as object-based rate control algorithms, description creation, or image and video quality evaluation.

1. INTRODUCTION

When working with image and video segmentation, the major objective is to design an algorithm that produces appropriate segmentation results for the particular goals of the application addressed. Nowadays, several applications exploit the representation of a video scene as a composition of video objects, taking advantage of the object-based standards for coding and representation specified by ISO: MPEG-4 [1] and MPEG-7 [2]. Examples are interactive applications that associate specific information and interactive "hooks" to the objects present in a given video scene, or applications that select different coding strategies, in terms of both techniques and parameter configurations, to encode the various video objects in the scene.

To enable such applications, the assessment of the image and video segmentation quality in view of the application goals assumes a crucial importance. In some cases, segmentation is automatically obtained using techniques like chroma-keying at the video production stage, but often the segmentation needs to be computed based on the image and video contents by using appropriate segmentation algorithms. Segmentation quality evaluation allows assessing the segmentation algorithm's adequacy for the targeted application, and it provides information that can be used to optimise the segmentation algorithm's behaviour by using the so-called relevance feedback mechanism [3].

Currently, there are no standard, or commonly accepted, methodologies available for objective evaluation of image or video segmentation quality. The current practice consists mostly in subjective ad hoc assessment by a representative group of human viewers. This is a time-consuming and expensive process for which no standard methodologies have been developed—often the standard subjective video quality evaluation guidelines are followed for test environment setup and scoring purposes [4, 5]. Nevertheless, efforts to propose objective evaluation methodologies and metrics have been intensified recently, with several proposals being available in the literature—see for instance [6–8].

Both subjective and objective segmentation quality evaluation methodologies usually consider two classes of evaluation procedures, depending on the availability, or not, of a reference segmentation taking the role of "ground truth," to be compared against the results of the segmentation algorithm under study. Evaluation against a reference is usually called relative, or discrepancy, evaluation, and when no reference is available it is usually called standalone, or goodness, evaluation.

Subjective evaluation, both relative and standalone, typically proceeds by analysing the segmentation quality of one object after another, with the human evaluators integrating the partial results and, finally, deciding on an overall segmentation quality score [9]. Objective evaluation automates all

the evaluation procedures, but the metrics available typically perform well only for very constrained applications scenarios [6].

Another distinction that is often made in terms of segmentation quality evaluation is if objects are taken individually, individual object evaluation, or if a segmentation partition[1] is evaluated, overall segmentation evaluation. The need for individual object segmentation quality evaluation is motivated by the fact that each video object may be independently stored in a database, or reused in a different context. An overall segmentation evaluation may determine, for instance, if the segmentation goals for a certain application have been globally met, and thus if a segmentation algorithm is appropriate for a given type of application. The evaluation of each object's relevance in the scene is essential for overall segmentation quality evaluation, as segmentation errors are less well tolerated for those objects that attract more the human visual attention.

This paper proposes metrics for the objective evaluation of video object relevance, namely, in view of objective overall segmentation quality evaluation. Section 2 presents the general methodology and metrics considered for overall video segmentation quality evaluation. The proposed methodology for video object relevance evaluation is presented in Section 3 and relevance evaluation metrics are proposed in Section 4. Results are presented in Section 5 and conclusions in Section 6.

2. OVERALL SEGMENTATION QUALITY EVALUATION METHODOLOGY AND METRICS

Both standalone and relative evaluation techniques can be employed for objective overall segmentation quality evaluation, whose goal is to produce an evaluation result for the whole partition. In this paper, the methodology for segmentation quality evaluation proposed in [6], including five main steps, is followed.

(1) *Segmentation*. The segmentation algorithm is applied to the test sequences selected as a representative of the application domain in question.

(2) *Individual object segmentation quality evaluation*. For each object, the corresponding individual object segmentation quality, either standalone or relative, is evaluated.

(3) *Object relevance evaluation*. The relevance of each object, in the context of the video scene being analyzed, is evaluated. Object relevance can be estimated by evaluating how much human visual attention the object is able to capture. Relevance evaluation is the main focus of this paper.

(4) *Similarity of objects evaluation*. The correctness of the match between the objects identified by the segmentation algorithm and those relevant to the targeted application is evaluated.

(5) *Overall segmentation quality evaluation*. The overall segmentation quality is evaluated by weighting the individual segmentation quality for the various objects in the scene with their relevance values, reflecting, for instance, the object's likeliness to be further reused or subject to some special processing that requires its shape to be as close as possible to the original. The overall evaluation also takes into account the similarity between the target set of objects and those identified by the segmentation algorithm.

The computation of the overall video segmentation quality metric (SQ) combines the individual object segmentation quality measures ($\mathrm{SQ\_io}_k$), for each object k, the object's relative contextual relevance ($\mathrm{RC\_rel}_k$), and the similarity of objects factor ($\mathrm{sim\_obj\_factor}$). To take into account the temporal dimension of video, the instantaneous segmentation quality of objects can be weighted by the corresponding instantaneous relevance and similarity of objects factors. The overall segmentation quality evaluation metric for a video sequence is expressed by

$$\mathrm{SQ} = \frac{1}{N} \cdot \sum_{t=1}^{N} \left[\mathrm{sim\_obj\_factor}_t \cdot \sum_{k=1}^{\mathrm{num\_objects}} \left(\mathrm{SQ\_io}_{kt} \cdot \mathrm{RC\_rel}_{k_t} \right) \right], \tag{1}$$

where N is the number of images of the video sequence, and the inner sum is performed for all the objects in the estimated partition at time instant t.

The individual object segmentation quality evaluation metric ($\mathrm{SQ\_io}_k$) differs for the standalone and relative cases. Standalone evaluation is based on the expected feature values computed for the selected object (intra-object metrics) and the disparity of some key features to its neighbours (inter-object metrics). The applicability and usefulness of standalone elementary metrics strongly depends on the targeted application and a single general-purpose metric is difficult to establish. Relative evaluation is based on dissimilarity metrics that compare the segmentation results estimated by the tested algorithm against the reference segmentation.

With the above overall video segmentation quality metric, the higher the individual object quality is for the most relevant objects, the better the resulting overall segmentation quality is, while an incorrect match between target and estimated objects also penalises segmentation quality.

3. VIDEO OBJECT RELEVANCE EVALUATION CONTEXT AND METHODOLOGY

Objective overall segmentation quality evaluation requires the availability of an object relevance evaluation metric, capable of measuring the object's ability to capture human visual attention. Such object relevance evaluation metric can also be useful for other purposes like description creation, rate control, or image and video quality evaluation. Object-based description creation can benefit from a relevance metric both directly as an object descriptor or as additional information. For instance, when storing the description of an

[1] A partition is understood as the set of non-overlapping objects that composes an image (or video frame), at a given time instant.

object in a database, the relevance measure can be used to select the appropriate level of detail for the description to store; more relevant objects should deserve more detailed and complete descriptions. Object-based rate control consists in finding and using, in an object-based video encoder, the optimal distribution of resources among the various objects composing a scene in order to maximise the perceived subjective image quality at the receiver. For this purpose, a metric capable of estimating in an objective and automatic way the subjective relevance of each of the objects to be coded is highly desirable, allowing a better allocation of the available resources. Also for frame-based video encoders, the knowledge of the more relevant image areas can be used to improve the rate control operation. In the field of image and video quality evaluation, the identification of the most relevant image areas can provide further information about the human perception of quality for the complete scene, thus improving image quality evaluation methodologies, as exemplified in [10].

The relevance of an object may be computed by considering the object on its own—individual object relevance evaluation—or adjusted to its context, since an object's relevance is conditioned by the simultaneous presence of other objects in the scene-contextual object relevance evaluation.

Individual object relevance evaluation

(RI) is of great interest whenever the object in question might be individually reused, as it gives an evaluation of the intrinsic subjective impact of that object. An example is an application where objects are described and stored in a database for later composition of new scenes.

Contextual object relevance evaluation

(RC) is useful whenever the context where the object is found is important. For instance, when establishing an overall segmentation quality measurement, or in a rate control scenario, the object's relevance to the context of the scene is the appropriate measure.

Both individual and contextual relevance evaluation metrics can be absolute or relative. *Absolute relevance metrics* (RI_abs and RC_abs) are normalised to the [0, 1] range, with value one corresponding to the highest relevance; each object can assume any relevance value independently of other objects. *Relative relevance metrics* (RI_rel and RC_rel) are obtained from the absolute relevance values by further normalisation, so that at any given instant the sum of the relative relevance values is one:

$$\mathrm{RC\_rel}_{kt} = \frac{\mathrm{RC\_abs}_{k_t}}{\sum_{j=1}^{\mathrm{num\_objects}} \mathrm{RC\_abs}_{j_t}}, \tag{2}$$

where $\mathrm{RC\_rel}_{kt}$ is the relative contextual object relevance metric for object k, at time instant t, which is computed from the corresponding absolute values for all objects (num_objects) in the scene at that instant.

The metrics considered for object relevance evaluation, both individual and contextual, are composite metrics involving the combination of several elementary metrics, each one capturing the effect of a feature that has impact on the object's relevance. The composite metrics proposed in this paper are computed for each time instant; the instantaneous values are then combined to output a single measurement for each object of a video sequence. This combination can be obtained by averaging, or taking the median of, the instantaneous values.

An object's relevance should reflect its importance in terms of human visual perception. Object relevance information can be gathered from various sources.

(i) *A priori information*. A way to rank object's relevance is by using the available a priori information about the type of application in question and the corresponding expected results. For instance, in a video-telephony application where the segmentation targets are the speaker and the background, it is known that the most important object is the speaking person. This type of information is very valuable, even if difficult to quantify in terms of a metric.

(ii) *User interaction*. Information on the relevance of each object can be provided through direct human intervention. This procedure is usually not very practical, as even when the objects in the scene remain the same, their relevance will often vary with the temporal evolution of the video sequence.

(iii) *Automatic measurement*. It is desirable to have an automatic way of determining the relevance for the objects present in a scene, at each time instant. The resulting measure should take into account the object's characteristics that make them instantaneously more or less important in terms of human visual perception and, in the case of contextual relevance evaluation, also the characteristics of the surrounding areas.

These three sources of relevance information are not mutually exclusive. When available, both a priori and user-supplied information should be used, with the automatic measurement process complementing them.

The methodology followed for the design of automatic evaluation video object relevance metrics consists in three main steps [11].

(1) *Human visual system attention mechanisms*. The first step is the identification of the image and video features that are considered more relevant for the human visual system (HVS) attention mechanisms, that is, the factors attracting viewers' attention (see Section 4.1).

(2) *Elementary metrics for object relevance*. The second step consists in the selection of a set of objective elementary metrics capable of measuring the relevance of each of the identified features (see Section 4.2).

(3) *Composite metrics for object relevance*. The final step is to propose composite metrics for individual and contextual video object's relevance evaluation, based on the elementary metrics above selected (see Section 4.3).

Ideally, the proposed metrics should produce relevance results that correctly match the corresponding subjective evaluation produced by human observers.

4. METRICS FOR VIDEO OBJECT RELEVANCE EVALUATION

Following the methodology proposed in Section 3, the human visual attention mechanisms are discussed in Section 4.1, elementary metrics that can be computed to automatically mimic the HVS behaviour are proposed in Section 4.2, and composite metrics for relevance evaluation are proposed in Section 4.3.

4.1. Human visual system attention mechanisms

The human visual attention mechanisms are determinant for setting up object relevance evaluation metrics. Objects that capture more the viewer's attention are those considered more relevant.

The HVS operates with a variable resolution, very high in the fovea and decreasing very fast towards the eye periphery. Directed eye movements (saccades) occur every 100–500 milliseconds to change the position of the fovea. Understanding the conditioning of these movements may help in establishing criteria for the evaluation of object relevance. Factors influencing eye movements and attention can be grouped into low-level and high-level factors, depending on the amount of semantic information they have associated.

Low-level factors influencing eye movements and viewing attention include the following [10].

(i) *Motion*. The peripheral vision mechanisms are very sensitive to changes in motion, this being one of the strongest factors in capturing attention. Objects exhibiting distinct motion properties from those of its neighbours usually get more attention.
(ii) *Position*. Attention is usually focused on the centre of the image for more than 25% of the time.
(iii) *Contrast*. Highly contrasted areas tend to capture more the viewing attention.
(IV) *Size*. Regions with large area tend to attract viewing attention; this effect, however, has a saturation point.
(V) *Shape*. Regions of long and thin shapes tend to capture more the viewer's attention.
(VI) *Orientation*. Some orientations (horizontal, vertical) seem to get more attention from the HVS.
(VII) *Colour*. Some colours tend to attract more the attention of human viewers; a typical example is the red colour.
(VIII) *Brightness*. Regions with high brightness (luminance) attract more attention.

High-level factors influencing eye movements and attention include the following [10].

(i) *Foreground/background*. Usually foreground objects get more attention than the background.
(ii) *People*. The presence of people, faces, eyes, mouth, hands usually attracts viewing attention due to their importance in the context of most applications.
(iii) *Viewing context*. Depending on the viewing context, different objects may assume different relevance values, for example, a car parked in a street or arriving at a gate with a car control.

Another important HVS characteristic is the existence of masking effects. Masking affects the perception of the various image components in the presence of each other and in the presence of noise [12]. Some image components may be masked due to noise (*noise masking*), similarly textured neighbouring objects may mask each other (*texture masking*), and the existence of a gaze point towards an object may mask the presence of other objects in an image (*object masking*). In terms of object relevance evaluation, texture and object masking assume a particular importance, since the simultaneous presence of various objects with different characteristics may lead to some of them receiving more attention than others.

4.2. Elementary metrics for object relevance evaluation

To automatically evaluate the relevance of an object, a number of elementary metrics are derived taking into account the human visual system characteristics. The proposal of the elementary relevance metrics should also take into account the previous work in this field; some relevant references are [10, 11, 13–16].

Each of the proposed elementary metrics is normalised to produce results in the [0, 1] range. Normalisation is done taking into account the dynamic range of each of the metrics, and in certain cases also by truncation to a range considered significant, determined after exhaustive testing with the MPEG-4 video test set.

The metrics considered are grouped, according to their semantic value, as low-level or high-level ones.

Low-level metrics

Both spatial and temporal features of the objects can be considered for computing low-level relevance metrics.

(1) *Motion activity*. This is one of the most important features according to the HVS characteristics. After performing global motion estimation and compensation to remove the influence of camera motion, two metrics that complement each other are computed.

(i) *Motion vectors average* (avg_mv) computes the sum of the absolute average motion vector components of the object at a given time instant, normalised by an image size factor:

$$\text{avg\_mv} = \frac{|\text{avg\_}X\text{\_vec}(k)| + |\text{avg\_}Y\text{\_vec}(k)|}{(\sqrt{\text{area}(I)/\text{area}(Q)}) \cdot 4}, \quad (3)$$

where avg_X_vec(k) and avg_Y_vec(k) are the average x and y motion vectors components for object k, area(I) is the image size and area(Q) is the size of a QCIF image (176 × 144). The result is truncated to the [0,1] range.

(ii) *Temporal perceptual information*(TI), proposed in [5] for video quality evaluation, is a measure of the amount of temporal change in a video. The TI metric

closely depends on the object differences for consecutive time instants, t and $t-1$:

$$\mathrm{TI}_{\mathrm{stdev}}(k_t) = \sqrt{\frac{1}{N}\cdot\sum_i\sum_j (k_t - k_{t-1})^2 - \left(\frac{1}{N}\cdot\sum_i\sum_j (k_t - k_{t-1})\right)^2}. \quad (4)$$

For normalisation purposes, the metric results are divided by 128 and truncated to the [0,1] range.

(2) *Size*. As large objects tend to capture more the visual attention, a metric based on the object's area, in pixels, is used. The complete image area is taken into account for normalisation of results:

$$\mathrm{size} = \begin{cases} 4\cdot\dfrac{\mathrm{area}(k)}{\mathrm{area}(I)}, & 4\cdot\mathrm{area}(k) < \mathrm{area}(I), \\ 1, & 4\cdot\mathrm{area}(k) \geq \mathrm{area}(I), \end{cases} \quad (5)$$

where k and I represent the object being evaluated and the image, respectively. It is assumed that objects covering, at least, one quarter of the image area are already large enough, thus justifying the inclusion of a saturation effect in this metric.

(3) *Shape and orientation*. The human visual system seems to prefer some specific types of shapes and orientations. Among these are long and thin, compact, and circular object shapes. Also horizontal and vertical orientations seem to be often preferred. A set of metrics to represent these features is considered: circularity (circ), elongation and compactness (elong_compact), and orientation (ori).

(i) *Circularity*. Circular-shaped objects are among the most preferred by human viewers and thus an appropriate metric of relevance is circularity:

$$\mathrm{circ}(k) = \frac{4\cdot\pi\cdot\mathrm{area}(k)}{\mathrm{perimeter}^2(k)}. \quad (6)$$

(ii) *Elongation and compactness*. A metric that captures the properties of elongation and compactness and combines them into a single measurement is proposed as follows:

$$\mathrm{elong\_compact}(k) = \frac{\mathrm{elong}(k)}{10} + \frac{\mathrm{compactness}(k)}{150}. \quad (7)$$

The weights in the formula were obtained after an exhaustive set of tests and are used for normalisation purposes together with a truncation at the limit values of 0 and 1.

Elongation can be defined as follows [17]:

$$\mathrm{elong}(k) = \frac{\mathrm{area}(k)}{(2\cdot\mathrm{thickness}(k))^2}, \quad (8)$$

where thickness(k) is the number of morphological erosion steps [18] that have to be applied to object k until it disappears.

Compactness is a measure of the spatial dispersion of the pixels composing an object; the lower the dispersion, the higher the compactness. It is defined as follows [17]:

$$\mathrm{compactness}(k) = \frac{\mathrm{perimeter}^2(k)}{\mathrm{area}(k)}, \quad (9)$$

where the perimeter is computed along the object border using a 4-neighbourhood.

(iii) *Orientation*. Horizontal and vertical orientations seem to be preferred by human viewers. A corresponding relevance metric is given by

$$\mathrm{orient} = \begin{cases} \left|3 - \dfrac{\mathrm{est\_ori}}{\pi/4}\right|, & \mathrm{est\_ori} > \frac{\pi}{2}, \\ \left|\dfrac{\mathrm{est\_ori}}{\pi/4} - 1\right|, & \mathrm{est\_ori} < \frac{\pi}{2}, \end{cases} \quad (10)$$

where est_ori is defined as [17]:

$$\mathrm{est\_ori} = \frac{1}{2}\cdot\tan^{-1}\left(\frac{2\cdot\mu_{11}(k)}{\mu_{20}(k)\cdot\mu_{02}(k)}\right), \quad (11)$$

with μ_{11}, μ_{02}, and μ_{20} being the first- and second-order centred moments for the spatial positions of the object pixels.

(4) *Brightness and redness*. Bright and coloured, especially red, objects seem to attract more the human visual attention. The proposed metric to evaluate these features is

$$\mathrm{brigh\_red} = \frac{3\cdot\mathrm{avg\_}Y(k) + \mathrm{avg\_}V(k)}{4\cdot 255}, \quad (12)$$

where $\mathrm{avg\_}Y(k)$ and $\mathrm{avg\_}V(k)$ compute the average values for the Y and V object colour components.

(5) *Object complexity*. An object with a more complex/detailed spatial content will usually tend to capture more attention. This fact can be measured using the spatial perceptual information (SI) and the criticality (critic) metrics for the estimated object.

(i) *Spatial perceptual information (*SI*)*. This is a measure of spatial detail, usually taking higher values for more (spatially) complex contents. It was proposed in [5] for video quality evaluation, based on the amplitude of the Sobel edge detector. SI can also be applied to an object k:

$$\mathrm{SI} = \max_{\mathrm{time}}\left(\mathrm{SI}_{\mathrm{stdev}}(k)\right) \quad (13)$$

with

$$\mathrm{SI}_{\mathrm{stdev}}(k) = \sqrt{\frac{1}{N}\cdot\sum_i\sum_j(\mathrm{Sobel}(k))^2 - \left(\frac{1}{N}\cdot\sum_i\sum_j(\mathrm{Sobel}(k))\right)^2}. \tag{14}$$

SI is normalised to the [0, 1] range dividing the metric results by 128, followed by truncation.

(ii) *Criticality* (critic). The criticality metric (crit) was proposed in [19] for video quality evaluation combining spatial and temporal information about the video sequence. For object relevance evaluation purposes, the proposed metric (critic) is applied to each object:

$$\mathrm{critic} = \frac{1-\mathrm{crit}}{5} \tag{15}$$

with

$$\begin{aligned}\mathrm{crit} &= 4.68 - 0.54\cdot p_1 - 0.46\cdot p_2,\\ p_1 &= \log_{10}\left(\mathrm{mean}_{\mathrm{time}}\left(\mathrm{SI}_{\mathrm{rms}}(k)\cdot\mathrm{TI}_{\mathrm{rms}}(k)\right)\right),\\ p_2 &= \log_{10}\left(\max_{\mathrm{time}}\left(\mathrm{abs}\left(\mathrm{SI}_{\mathrm{rms}}(k_t) - \mathrm{SI}_{\mathrm{rms}}(k_{t-1})\right)\right)\right)\\ \mathrm{SI}_{\mathrm{rms}}(k) &= \sqrt{\frac{1}{N}\cdot\sum_i\sum_j(\mathrm{Sobel}(k))^2},\\ \mathrm{TI}_{\mathrm{rms}}(k_t) &= \sqrt{\frac{1}{N}\cdot\sum_i\sum_j(k_t - k_{t-1})^2}.\end{aligned} \tag{16}$$

(6) *Position*. Position is an important metric for contextual evaluation, as the fovea is usually directed to the centre of the image around 25% of the time [10]. The distance of the centre of gravity of object k to the image (I) centre is used as the position metric:

$$\mathrm{pos} = 1 - \frac{\left|\mathrm{grav\_}Xc(I) - \mathrm{grav\_}Xc(k)\right|/\mathrm{grav\_}Xc(I) + \left|\mathrm{grav\_}Yc(I) - \mathrm{grav\_}Yc(k)\right|/\mathrm{grav\_}Yc(I)}{2}, \tag{17}$$

where $\mathrm{grav\_}Xc(k)$ and $\mathrm{grav\_}Yc(k)$ represent, respectively, the x-and y-coordinates of the centre of gravity of object k. The normalisation to the [0,1] range is guaranteed by truncation.

(7) *Contrast to neighbours*. An object exhibiting high contrast values to its neighbours tends to capture more the viewer attention, thus being more relevant. The metric proposed for its evaluation measures the average maximum local contrast of each pixel to its neighbours at a given time instant:

$$\mathrm{contrast} = \frac{1}{4\cdot N_b}\cdot\sum_{i,j}\left(2\cdot\max\left(DY_{ij}\right) + \max\left(DU_{ij}\right) + \max\left(DV_{ij}\right)\right), \tag{18}$$

where N_b is the number of border pixels of the object, and DY_{ij}, DU_{ij}, and DV_{ij} are measured as the differences between an object's border pixel, with Y, U, and V components, and its 4-neighbours.

Notice that the position and contrast metrics are applicable only for contextual relevance evaluation.

High-level metrics

These are metrics involving some kind of semantic understanding of the scene.

(1) *Background*. whether an object belongs to the background or to the foreground of a scene influences the user attention devoted to that object, with foreground objects typically receiving a larger amount of attention. Additionally, it is possible to distinguish the various foreground objects according to their depth levels. Typically, objects moving in front of other objects receive a larger amount of visual attention.

A contextual relevance metric, called background, may be associated to this characteristic of an object, taking a value between zero (objects belonging to the background) and one (topmost foreground objects). Desirably, depth estimation can be computed using automatic algorithms, eventually complemented with user assistance to guarantee the desired meaningfulness of the results. User input may be provided when selecting the object masks corresponding to each object, for example, by checking a background flag in the dialog box used.

The proposed background metric is

$$\mathrm{background} = \begin{cases}0, & n = 0,\\ 0.5\cdot\left(1 + \frac{n}{N}\right), & n \neq 0,\end{cases} \tag{19}$$

where n takes value 0 for the background components, and a depth level ranging from 1 to N for the foreground objects. The highest value is attributed to the topmost foreground object. This metric distinguishes the background from the foreground objects, thus receiving the name background, even if a distinction between the various foreground objects according to their depth is also performed.

(2) *Type of object*. Some types of objects usually get more attention from the user due to their intrinsic semantic value. For instance, when a person is present in an image it usually gets high viewer attention, in particular the face area. Or, for

an application that automatically reads car license plates, the most relevant objects are the cars and their license plates. If algorithms for detecting the application-relevant objects are available, their results can provide useful information for object relevance determination. In such cases, the corresponding metric would take value one when a positive detection occurs and zero otherwise.

Apart from the metrics that explicitly include information about the context where the object is identified (position, contrast to neighbours and background), which make sense only for contextual relevance evaluation, the remaining metrics presented can be considered for both individual and contextual relevance evaluation.

4.3. Composite metrics for object relevance evaluation

This section proposes composite metrics for individual and for contextual object relevance evaluation. As different sequences present different characteristics, a single elementary metric, which is often related to a single HVS property, is not expected to always adequately estimate object relevance. This leads to the definition of *composite metrics* that integrate the various factors to which the HVS is sensitive to be able to provide robust relevance results independently of the particular segmentation partition under consideration.

The combination of elementary metrics into composite ones was done after an exhaustive set of tests, using the MPEG-4 test set, with each elementary metric behaviour being subjectively evaluated by human observers.

For individual relevance, only an absolute metric is proposed, providing relevance values in the range [0,1]. For contextual relevance, the objective is to propose a relative metric to be used in segmentation quality evaluation, providing object relevance values that, at any temporal instant, sum to one. These relative contextual relevance values are obtained from the absolute contextual relevance values by using (2). To obtain a relevance evaluation representative of a complete sequence or shot, a temporal integration of the instantaneous values can be done by performing a temporal average or median of the instantaneous relevance values.

Composite metric for individual object relevance evaluation

The selection of weights for the various elementary relevance metrics is done taking into account the impact of each metric in terms of its ability to capture the human visual attention, complemented by each elementary metric's behaviour in the set of tests performed. The result was the assignment of the largest weights to the motion activity and complexity metrics. The exact values selected for the weights of the various classes of metrics, and for the elementary metrics within each class represented by more than one elementary metrics, resulted from an exhaustive set of tests. It is worth recalling that for individual relevance evaluation, the elementary metrics of position, contrast and background cannot be used.

The proposed composite metric for *absolute individual object relevance* evaluation ($\mathrm{RI\_abs}_k$) for an object k, which produces relevance values in the range [0,1], is given by

$$\mathrm{RI\_abs}_k = \frac{1}{N} \cdot \sum_{t=1}^{N} \mathrm{RI\_abs}_{kt}, \tag{20}$$

where N is the total number of temporal instances in the segmented sequence being evaluated, and the instantaneous values of $\mathrm{RI\_abs}_{kt}$ are given by

$$\begin{aligned} &\mathrm{RI\_abs}_{kt} \\ &\quad = 0.38 \cdot \mathrm{mot\_activ}_t + 0.33 \cdot \mathrm{comp}_t + 0.14 \cdot \mathrm{shape}_t \\ &\qquad + 0.1 \cdot \mathrm{bright\_red}_t + 0.05 \cdot \mathrm{size}_t \end{aligned} \tag{21}$$

with

$$\begin{aligned} \mathrm{mot\_activ}_t &= 0.57 \cdot \mathrm{avg\_mv}_t + 0.43 \cdot \mathrm{TI}_t \\ \mathrm{shape}_t &= 0.4 \cdot \mathrm{circ}_t + 0.6 \cdot \mathrm{elong\_compact}_t \\ \mathrm{comp}_t &= 0.5 \cdot \mathrm{SI}_t + 0.5 \cdot \mathrm{critic}_t . \end{aligned} \tag{22}$$

The instantaneous values of the *relative individual object relevance* evaluation ($\mathrm{RI\_rel}_{kt}$) can be obtained from the corresponding absolute individual relevance ($\mathrm{RI\_abs}_{ki}$) metric by applying (2).

Composite metric for contextual object relevance evaluation

The composite metric for absolute contextual object relevance evaluation ($\mathrm{RC\_abs}_k$) produces relevance values between 0 and 1. Its main difference regarding the absolute individual object relevance metric ($\mathrm{RI\_abs}_k$) is that the contextual elementary metrics can now be additionally taken into account.

The proposed metric for the instantaneous values of the *absolute contextual object relevance* ($\mathrm{RC\_abs}_{kt}$) is given by

$$\begin{aligned} &\mathrm{RC\_abs}_{kt} \\ &\quad = 0.3 \cdot \mathrm{motion\_activ}_t + 0.25 \cdot \mathrm{comp}_t + 0.13 \cdot \mathrm{high\_level}_t \\ &\qquad + 0.1 \cdot \mathrm{shape}_t + 0.085 \cdot \mathrm{bright\_red}_t + 0.045 \\ &\qquad \cdot (\mathrm{contrast}_t + \mathrm{position}_t + \mathrm{size}_t), \end{aligned} \tag{23}$$

with $\mathrm{motion\_activ}_i$, shape_i, and comp_i defined as for the $\mathrm{RI\_abs}_k$ composite metric, and $\mathrm{high\_level}_t$ defined as

$$\mathrm{high\_level}_t = \mathrm{background}_t . \tag{24}$$

The proposed metric for computing the instantaneous values of the *relative contextual object relevance* evaluation ($\mathrm{RC\_rel}_{kt}$), which produces a set of relevance values that sum to one at any time instant, is obtained from the corresponding absolute contextual relevance ($\mathrm{RC\_abs}_{ki}$) metric by applying (2).

Finally, the relative contextual object relevance evaluation metric ($\mathrm{RC\_rel}_k$) producing results for the complete duration of the sequence is given by the temporal average of the instantaneous values:

$$\mathrm{RC\_rel}_k = \frac{1}{N} \cdot \sum_{t=1}^{N} \mathrm{RC\_rel}_{kt}. \tag{25}$$

(a) (b) (c) (d)

FIGURE 1: Sample frames of the test sequences: Akiyo (a), Hall Monitor (b), Coastguard (c) and Stefan (d).

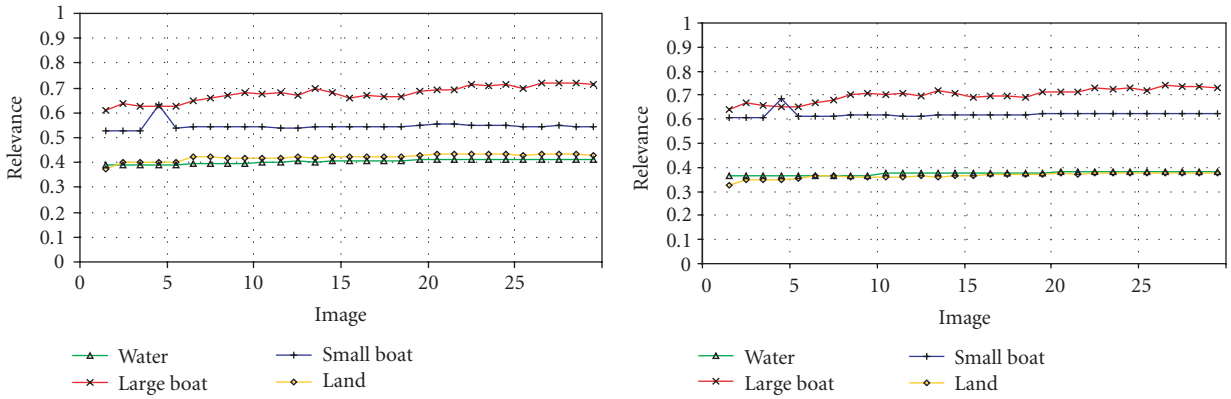


FIGURE 2: Individual and contextual absolute relevance metrics for a portion of the Coastguard sequence.

The relevance evaluation algorithm developed is completely automatic as far as the low-level metrics are concerned. The only interaction requested from the user in terms of contextual relevance evaluation regards the classification of objects as background or foreground, and eventually the identification of the depth levels for the foreground objects (if this is not done automatically).

5. OBJECT RELEVANCE EVALUATION RESULTS

Since this paper is focused on object relevance evaluation for objective evaluation of overall segmentation quality, the most interesting set of results for this purpose are those of relative contextual object relevance evaluation. However, for completeness, also individual object relevance results are included in this section. The object relevance results presented here use the MPEG-4 test sequences "Akiyo," "Hall Monitor," "Coastguard," and "Stefan," for which sample frames are included in Figure 1. The objects for which relevance is estimated are obtained from the corresponding reference segmentation masks available from the MPEG-4 test set, namely: "Newsreader" and "Background" for sequence "Akiyo;" "Walking Man" and "Background" for sequence "Hall Monitor;" "Tennis Player" and "Background" for sequence "Stefan;" "Small Boat," "Large Boat," "Water," and "Land" for sequence "Coastguard."

Examples of absolute relevance evaluation results are included in Figures 2 and 3. These figures show the temporal evolution of the instantaneous absolute individual and contextual relevance values estimated for each object, in samples of the Coastguard and Stefan sequences.

Figure 4 shows a visual representation of each object's temporal average of absolute contextual object relevance values, where the brighter the object is, the higher its relevance is.

Examples of relative object relevance results are provided in Table 1. The table includes the temporal average values of both the individual (Indiv) and contextual (Context) relative object relevancies, computed using the proposed metrics for each object of the tested sequences.

Individual object relevance results show that objects with larger motion activity and more detailed spatial content tend to achieve higher metric values. For instance, the background object in the Akiyo sequence gets the lowest absolute individual relevance value (RI_abs = 0.23, RI_rel = 0.36), as it is static and with a reasonably uniform spatial content. On the other hand, the tennis player object of the Stefan sequence is considered the most relevant object (RI_abs = 0.73, RI_rel = 0.58), mainly because it includes a considerable amount of motion.

Contextual object relevance results additionally consider metrics such as the *spatial position* of the object, its contrast

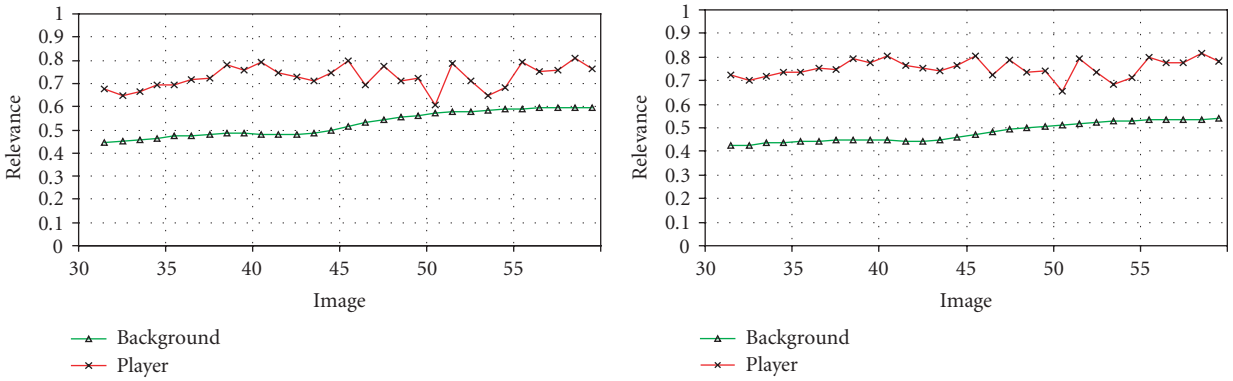
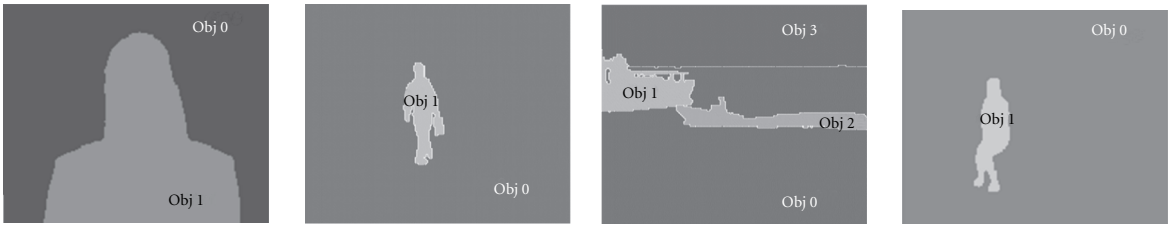


Figure 3: Individual and contextual absolute relevance metrics for a portion of the Stefan sequence.



(a) (b) (c) (d)

Figure 4: Visual representation of each object's temporal average of absolute contextual object relevance values for the Akiyo (a), Hall Monitor (b), Coastguard (c) and Stefan (d) sequences.

to the neighbours and the information about belonging or not to the background, which have an important role in terms of the HVS behaviour. Comparing the individual and contextual relative relevance values, included in Table 1, for instance for the Stefan sequence, it is possible to observe that the relative individual object relevancies are 0.42 and 0.58 for the background and tennis player objects, respectively, while the corresponding contextual values are 0.39 and 0.61. These results show that by using the additional contextual elementary metrics the tennis player gets a higher relevance value, as could be expected from a subjective evaluation.

To support the above conclusion, a set of informal subjective tests was performed. These tests were performed by a restricted number of test subjects (ten), mainly people working at the Telecommunications Institute of Instituto Superior Técnico, Lisbon, Portugal. The test subjects were shown the various test sequences as well as the various segmented objects composing each partition, over a grey background, and were asked to give an absolute contextual object relevance score for each object in the [0,1] range; these absolute scores were then converted into relative scores using (2). Relevance was defined to the test subjects as the ability of the object to capture the viewer attention. Table 1 also includes the average subjective test results (Subj) together with their differences (Diff) from the relative contextual object relevance values computed automatically (Obj).

These results show a close match between the objective/automatic object relevance evaluation and the informal subjective tests. The only significant differences occur for the two sequences containing "human objects," notably people facing the camera. In this case, the automatic algorithms underestimated the corresponding object relevance values. This observation reinforces the need for inclusion, whenever available, of the high-level *type of object* metric, namely, to appropriately take into account the presence of people.

Another difference can be observed in the results for the Coastguard sequence, where the automatic classification system gave higher relevance values to the large boat, while test subjects ranked it as equally relevant to the small boat. In this case, the fact that the camera was following the small boat had a large impact on the subjective results, while the automatic metrics only partially captured the HVS behaviour. To better cover this case, the motion activity class of metrics could take into account not only the motion of the object but also its relation to the camera motion.

In general, the automatically computed results presented above tend to agree with the human subjective impression of the object's relevance. It can be noticed that for all the tested cases, the objects have been adequately ranked by the composite objective relevance evaluation metrics. Contextual metrics tend to agree better with the subjective assessment of relevance, which typically takes into account the

Table 1: Temporal average of objective individual (Indiv) and contextual (Context-Obj) relative relevance values for each object of the test sequences considered. For contextual relevance values, the average subjective (Subj) values obtained from a limited subjective evaluation test and the corresponding differences (Diff) from automatically computed values are also included.

Akiyo	Background (Obj **0**)				Newsreader (Obj **1**)			
	Indiv	Context			Indiv	Context		
		Obj	Subj	Diff		Obj	Subj	Diff
	0.36	0.33	0.25	−0.08	0.64	0.67	0.75	0.08
Hall Monitor	Background (Obj **0**)				Walking man (Obj **1**)			
	Indiv	Context			Indiv	Context		
		Obj	Subj	Diff		Obj	Subj	Diff
	0.38	0.36	0.34	−0.02	0.62	0.64	0.66	0.02
Stefan	Background (Obj **0**)				Tennis player (Obj **1**)			
	Indiv	Context			Indiv	Context		
		Obj	Subj	Diff		Obj	Subj	Diff
	0.42	0.39	0.35	−0.04	0.58	0.61	0.65	0.04
Coastguard	Water (Obj **0**)				Large boat (Obj **1**)			
	Indiv	Context			Indiv	Context		
		Obj	Subj	Diff		Obj	Subj	Diff
	0.20	0.18	0.12	−0.06	0.33	0.34	0.36	0.02
	Small boat (Obj **2**)				Land (Obj **3**)			
	Indiv	Context			Indiv	Context		
		Obj	Subj	Diff		Obj	Subj	Diff
	0.27	0.30	0.36	0.06	0.20	0.18	0.16	−0.02

context where the object is found. Even when the context of the scene is not considered, the absolute individual object relevance metrics (not using the position, contrast, and background metrics) manage to successfully assign higher relevance values to those objects that present characteristics that attract most the human visual attention.

6. CONCLUSIONS

The results obtained with the proposed object relevance evaluation metrics indicate that an appropriate combination of elementary metrics, mimicking the human visual system attention mechanisms behaviour, makes it possible to have an automatic system to automatically measure the relevance of each video object in a scene. This paper has proposed contextual and individual object relevance metrics, applicable whenever the object context in the scene should, or should not, be taken into account, respectively. In both cases, absolute and relative relevance values can be computed.

For overall segmentation quality evaluation, the objective metric to be used is the relative contextual object relevance, as it expresses the object's relevance in the context of the scene. This is also the metric to be used for rate control or image quality evaluation scenarios, as discussed in Section 3. From the results in Section 5, it was observed that the proposed objective metric for relative contextual object relevance achieves results in close agreement with the subjective relevance perceived by human observers. As an example, a mobile video application that segments the video scene into a set of objects can be considered. This application would make use of the relative contextual relevance metric to select for transmission only the most relevant objects and allocate the available coding resources among these objects according to their instantaneous relevancies.

The absolute individual object relevance metric can also play an important role in applications such as description creation. An example is the management of a database of video objects that are used for the composition of new video scenes using the stored objects. In this type of application, objects can be obtained from the segmentation of natural video sequences and stored in the database together with descriptive information. The objects to be stored in the database as well as the amount of descriptive information about them can be decided taking into consideration the corresponding relevancies.

REFERENCES

[1] ISO/IEC 14496, "Information technology—Coding of Audio-Visual Objects," 1999.

[2] ISO/IEC 15938, "Multimedia Content Description Interface," 2001.

[3] Y. Rui, T. S. Huang, and S. Mehrotra, "Relevance feedback techniques in interactive content-based image retrieval," in *Proceedings of IS&T SPIE Storage and Retrieval for Image and Video Databases VI*, vol. 3312 of *Proceedings of SPIE*, pp. 25–36, San Jose, Calif, USA, January 1998.

[4] ITU-R, "Methodology for the Subjective Assessment of the Quality of Television Pictures," Recommendation BT.500-7, 1995.

[5] ITU-T, "Subjective Video Quality Assessment Methods for Multimedia Applications," Recommendation P.910, August 1996.

[6] P. L. Correia and F. Pereira, "Objective evaluation of video segmentation quality," *IEEE Transactions on Image Processing*, vol. 12, no. 2, pp. 186–200, 2003.
[7] C. E. Erdem, A. M. Tekalp, and B. Sankur, "Metrics for performance evaluation of video object segmentation and tracking without ground-truth," in *Proceedings of IEEE International Conference on Image Processing (ICIP '01)*, vol. 2, pp. 69–72, Thessaloniki, Greece, October 2001.
[8] P. Villegas and X. Marichal, "Perceptually-weighted evaluation criteria for segmentation masks in video sequences," *IEEE Transactions on Image Processing*, vol. 13, no. 8, pp. 1092–1103, 2004.
[9] COST211quat European Project, "Call for AM Comparisons," available at: http://www.iva.cs.tut.fi/COST211/Call/Call.htm.
[10] W. Osberger, N. Bergmann, and A. Maeder, "A technique for image quality assessment based on a human visual system model," in *Proceedings of 9th European Signal Processing Conference (EUSIPCO '98)*, pp. 1049–1052, Rhodes, Greece, September 1998.
[11] P. L. Correia and F. Pereira, "Estimation of video object's relevance," in *Proceedings of 10th European Signal Processing Conference (EUSIPCO '00)*, pp. 925–928, Tampere, Finland, September 2000.
[12] T. Hamada, S. Miyaji, and S. Matsumoto, "Picture quality assessment system by three-layered bottom-up noise weighting considering human visual perception," in *Proceedings of 139th SMPTE Technical Conference*, pp. 179–192, New York, NY, USA, November 1997.
[13] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
[14] X. Marichal, T. Delmot, C. Vleeschouwer, V. Warscotte, and B. Macq, "Automatic detection of interest areas of an image or of a sequence of images," in *Proceedings of IEEE International Conference on Image Processing (ICIP '96)*, vol. 3, pp. 371–374, Lausanne, Switzerland, September 1996.
[15] F. W. M. Stentiford, "An estimator for visual attention through competitive novelty with application to image compression," in *Proceedings of Picture Coding Symposium*, pp. 101–104, Seoul, Korea, April 2001.
[16] J. Zhao, Y. Shimazu, K. Ohta, R. Hayasaka, and Y. Matsushita, "A JPEG codec adaptive to region importance," in *Proceedings of 4th ACM International Conference on Multimedia (ACM Multimedia '96)*, pp. 209–218, Boston, Mass, USA, November 1996.
[17] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis and Machine Vision*, Chapman & Hall, London, UK, 1993.
[18] J. Serra, *Image Analysis and Mathematical Morphology*, vol. 1, Academic Press, London, UK, 1993.
[19] S. Wolf and A. Webster, "Subjective and objective measures of scene criticality," in *Proceedings of ITU Meeting on Subjective and Objective Audiovisual Quality Assessment Methods*, Turin, Italy, October 1997.

Paulo Correia graduated as an engineer and obtained the M.S. and Ph.D. degrees in electrical and computers engineering from Instituto Superior Técnico (IST), Technical University of Lisbon , Portugal, in 1989, 1993, and 2002, respectively. He is currently Assistant Professor at the Electrical and Computer Engineering Department of IST, and since 1994 he has been a researcher at the Image Group of Instituto de Telecomunicações. He has participated in several national and international research projects, being the National Representative of the the European COST 292 project. He is a member of the EURASIP/Elsevier Signal Processing Journal editorial board. He is an Elected Member of the EURASIP Administrative Committee. Current research interests include video analysis and processing, namely, video segmentation, objective video segmentation quality evaluation, content-based video description, and biometric recognition.

Fernando Pereira was born in Vermelha, Portugal in October 1962. He graduated from the Electrical and Computer Engineering Department, Instituto Superior Técnico (IST), Technical University of Lisbon, Portugal, in 1985. He received the M.S. and Ph.D. degrees in electrical and computer engineering from IST in 1988 and 1991, respectively. He is currently Professor at the Electrical and Computer Engineering Department of IST. He is responsible for the participation of IST in many national and international research projects. He acts often as project evaluator and auditor for various organisations. He is a Member of the Editorial Board and Area Editor on image/video compression of signal processing of Image Communication Journal, a Member of the IEEE Press Board, and an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Image Processing, and IEEE Transactions on Multimedia. He is an IEEE Distinguished Lecturer and a member of the scientific and program committees of tens of international conferences and workshops. He has contributed more than 150 papers to journals and international conferences. He won the 1990 Portuguese IBM Award and an ISO Award for Outstanding Technical Contribution for his participation in the development of the MPEG-4 Visual standard. He has been participating in the work of ISO/MPEG for many years, notably as the Head of the Portuguese Delegation, Chairman of the MPEG Requirements Group, and chairing many ad hoc groups related to the MPEG-4 and MPEG-7 standards. His current areas of interest are video analysis, processing, coding and description, and multimedia interactive services.

Composition Comments

1. We moved "both" to its current place and added the highlighted "and." Please check.

2. Should we change "detail" to "details"? Please check.

3. We changed "in" to "to." Please check.

4. We added "in the presence of." Please check.

5. We changed the "x" to the highlighted "$\times$." Please check.

6. We changed "reminding" to "recalling." Please check.

7. Should we change the subscript "i" in the highlighted "$motion\_activ_i$," "$shape_i$," and "$comp_i$" to "t"? Pleas check.

8. Can we change "with" to "it has"? Please check.

9. We changed "to" to "from." Please check highlighted cases throughout.

10. Comment on ref. [16]: We changed the name of the 3rd author. Please check.

11. Comment on ref. [18]: Should we change the year from "1993" to "1982". Please check.

12. Please note that the biographies should not exceed 200 words. Consequently, please reduce the biography of "Fernando Pereira" to the required number of words.

13. We added the highlighted "Department." Please check.