A TRIPLE USER CHARACTERIZATION MODEL FOR VIDEO ADAPTATION AND QUALITY OF EXPERIENCE EVALUATION

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

1. INTRODUCTION

Video services, especially those involving communications, engage technology with its associated limitations as well as the human users who also have associated limitations, or more generally speaking characteristics and preferences. In this context, the service goal is typically maximizing 'quality of service' for the available resources or minimizing the required resources for a prescribed quality of service. The growing heterogeneity of networks, terminals and users and the increasing availability and usage of multimedia content have been raising the relevance of content adaptation technologies able to fulfill the needs associated to all usage conditions without multiplying the number of versions available for the same piece of content while simultaneously maximizing user satisfaction. The notion of Universal Multimedia Access (UMA) calls for the provision of different presentations of the same information, with more or less complexity, suiting different usage environments (i.e., the context) in which the content will be consumed; for this purpose, multimedia content has to be adapted either off-line or in real-time.

While universal multimedia adaptation is still in its infancy it has already become clear that, as delivery technology evolves, the human factors associated with multimedia consumption assume an increasing importance. In particular, the importance of the user rather than the terminal as the final point in the multimedia consumption chain is becoming clear. We are starting to speak about Universal Multimedia Experiences (UME) which provide the users with adapted, informative (in the sense of cognition), and exciting (in the sense of feelings) experiences [1]. Following the same trends, the notion of 'quality of service' has to evolve to something more encompassing like 'quality of experience' where user satisfaction considers not only the sensorial and perceptual dimensions but also the important emotional dimension.

Sensations, perceptions and emotions play a central role in major multimedia applications such as video streaming, personal communications, and image and music libraries, determining final user satisfaction. In this context, this paper proposes the triple sensation-perception-emotion user characterization model for content adaptation and discusses adaptation tools and quality of experience metrics in light of this model.

Multimedia content adaptation is a central topic for the MPEG-7 and MPEG-21 standards. While MPEG-7 provides content description tools, MPEG-21 provides context (or usage environment) description tools as well as other useful tools for content adaptation. However, the decision mechanisms regarding which adaptations to perform and the algorithms to perform them are outside the scope of these standards as well as the evaluation of the resulting user experience.

In this context, this paper is complementary to the MPEG-7 and MPEG-21 standards. The major purpose of this paper is to launch some new ideas in the field of video adaptation and quality of experience evaluation. It is well recognized that there is much work to do in order these ideas are better studied and eventually validated.

2. ABOUT SENSATIONS, PERCEPTIONS AND EMOTIONS

Without going deep in the intricacies of the theories of perception and behavior, it is generally considered that the perceptual process consists of two major stages: sensation and perception [2]. Sensations regard the first contact between the human organism and the environment; sensations regard the simple conscious experience associated to a stimulus, for example light; sensations regard the faculty through which the external world is apprehended. The characteristics of the eye, the human optical system, determine the way sensations are created. Perceptions regard the conscious experience of objects and object relationships; perceptions regard the identification of objects. While sensations are clearly monomodal, perceptions may be multimodal: the question, "Can you identify that object?" may be answered using only visual sensations, combining visual and aural sensations, or making any other combination of sensations.

While the line between sensations and perceptions is not fully sharp since there are no perceptions without sensations, it seems clear that sensations are more low-level, and less related to the composition of the real world than perceptions. On the other hand, perceptions are about the human user becoming aware of something using the senses and the brain, this means about the so called 'what' and 'where'. Since perceptions regard objects and their location, and these objects have a purpose or a meaning, perception is deeply related to semantics that is to the human meaning of things. After perception, sensations are not anymore only an organic reaction to a stimulus but a structured representation of the surrounding world. In general, the so-called human visual system (HVS) includes both the sensorial and perceptual processes.

While perception is part of the human cognition process (this means the process of knowing, which also includes learning), human behavior also includes affections or emotions which are psychic and physical reactions (as joy, anger or fear) subjectively experienced as strong feelings and physiologically involving changes that may prepare the body for immediate reactions. Emotions play a central role in human life, behavior and relations; emotions play a central role in communication as well as in entertainment which means that any consumption of visual information involves sensations and perceptions and the associated mechanisms but also emotions and feelings.

3. THE TRIPLE USER MODEL FOR VIDEO ADAPTATION AND EVALUATION

Since sensations, perceptions and emotions play a major role in the consumption of multimedia content, this paper proposes a triple sensation-perception-emotion user characterization model for the evaluation of the quality of experience, notably when video adaptation is performed. This model is hierarchical in the sense that typically emotions build on perceptions and perceptions build on sensations, setting a hierarchy.

While the current vision of the video adaptation process sees it mostly conditioned by the resources available, especially in terms of networks and devices, this is not always the case since the maximization of user satisfaction may require some adaptation processing even if there are no resource constraints. Here the driving force for adaptation would not be the 'resource constraints' part of the equation but the 'satisfaction maximization' part of it. This is a rather important conceptual jump in the area of video adaptation which is absent from most of the relevant research published in the area.

As will be seen in the following, a major issue in video adaptation and in the usage of the triple user model proposed is the dimensionality of the adaptation solution in terms of modalities. While most video adaptation solutions in the literature are monomodal (also called transcoding), this means video is adapted to video, this may be a rather conservative approach considering that users consume multimedia content with more modalities and senses. So, multimodal adaptation solutions (also called transmoding) are clearly more powerful in 'filling' the human senses and thus have a higher probability of reaching higher quality multimedia experiences. Both for transcoding and transmoding, adaptation may be implemented together with some higher-level semantic filtering related to the user preferences. This semantic filtering works clearly at the emotional level for most entertainment services, e.g., video summary according to some criteria.

Transcoding and transmoding solutions are different not only at the emotional level but also at the sensorial and perceptual levels. This fact implies that the evaluation of video adaptation solutions, monomodal or transmodal, has to consider at least three quality evaluation dimensions directly related to the triple user model layers:

- **Sensorial evaluation -** This level of user experience evaluation regards the user satisfaction in terms of sensorial experience; for video, this means content sharpness, blurriness, brightness, naturalness, lack of artifacts, etc. independently of what the content contains. Although, the user experience always implies perception and in that sense there are no sensorial experiences without perceptual experiences (so there is formally no sensorial evaluation but always perceptual evaluation), we will define sensorial evaluation the evaluation that simply regards the sensation of 'looking better' independently of what is in the content; it is acknowledged that this type of evaluation is often called perceptual evaluation in the literature but we reserve the term perceptual evaluation for another concept.
- **Perceptual evaluation –** This level of user experience evaluation regards the user satisfaction in terms of perceptual or cognitive experience; this means in terms of the amount of knowledge the user acquires about what is in the content, where, etc. For example, in a transmoding process, the sensorial quality may be very high but the perceptual quality may be low if the amount of information provided is significantly reduced, e.g., because the wrong shots have been chosen for the summary.
- **Emotional evaluation -** This level of user experience evaluation regards the user satisfaction in terms of emotional experience; this means in terms of the intensity of the feelings the user experiences. The adaptation may target 'good' or 'bad' emotions, for example fear is targeted in horror content, and thus the evaluation must check the efficacy in targeting the right emotion; for example, it would not be adequate (in principle) that an adaptation aiming at happiness causes sadness or anger.

These types of evaluation have to be supported by adequate evaluation methodologies, either subjective or objective. Since humans are the ultimate receivers for most relevant applications, it is clear that the most reliable way of assessing the sensorial, perceptual and emotional qualities of a video experience is through subjective evaluation. Both for subjective and objective evaluation, the methods and the metrics must be related to the three layers of evaluation. Since the triple user model is hierarchical, the sensorial quality impacts on the perceptual quality and the perceptual quality on the emotional quality. In this context, all possible video adaptation solutions are driven by some sensorial, perceptual or emotional factors or any combination of them.

The definition of the most adequate type of adaptation may be based on the so called utility theory which defines the adaptation problem as one of maximizing the utility for the relevant resource constraints. In the context of this paper, the utility would be associated to the sensorial, perceptual and emotional dimensions of the quality of experience. The adequate usage by the adaptation mechanism of available user preferences data will imply higher quality perceptual and emotional experiences.

3.1 Adaptation and quality at the sensorial layer

Most of the transcoding adaptation solutions work at the sensorial level this means transcoding is performed trying to provide the user a better sensorial experience under the existing resources constraints in order a better perceptual experience may also happen after. Examples of this type of adaptation are transcoding at the level of:

- **Component –** e.g., color to black and white**.**
- **Color depth –** e.g., 8 to 4 bit/sample.
- **Bitrate or quality –** e.g., 256 to 32 kbit/s.
- **Error resilience –** e.g., low to high resilience.
- **Spatial resolution –** e.g., CIF to QCIF.
- **Temporal resolution –** e.g., 25 to 12,5 Hz.
- **Bitstream statistics –** e.g., CBR to VBR.
- **Visual handicaps –** e.g., pixel-based adaptation for color blind deficiencies.
- **Natural environment –** e.g., pixel-based adaptation for low illumination environments.

For most cases, sensorial level quality evaluation regards the similarity or fidelity of the adapted version with the non adapted version of the content. This is not necessarily the case for visual handicaps adaptation where video data is intentionally distorted to produce a better sensorial experience. In general, it may be said that sensorial evaluation is about measuring the 'looking nice' feeling stimulated at the user.

Subjective assessment is clearly the most reliable way of performing sensorial quality evaluation and there is a set of methodologies already defined, and largely used, in the development and verification of video coding standards, e.g., the Double Stimulus and Single Stimulus methods. However, because these methodologies are inconvenient, slow and expensive, there has been a great deal of research on objective quality assessment targeting the design of metrics that can reliably and automatically predict video quality. Of course, the mean squared error (MSE) and the peak signal-to-noise-ratio (PSNR) are the most popular objective video quality metrics but it is well know that they are limited in their capacity to replicate (and correlate with) subjective assessments, e.g. in terms of mean opinion score (MOS). While there are already many objective video quality assessment metrics available in the literature (mainly full-reference), this is still mostly an open problem.

Regarding the quality evaluation of sensorial-based adaptations, the available subjective and objective video quality assessment methods may be used; here the coded but not adapted version of the content plays the role of the original, undistorted version while the adapted version plays the role of the distorted version. This approach may not work when using the available objective metrics for visual handicaps or natural environment adaptation; for this case, subjective evaluation may well be the only solution.

3.2 Adaptation and quality at the perceptual layer

Adaptation at the perceptual level may involve either transcoding or transmoding always with the target to maximize the cognitive experience regarding the world represented by the content at hand, this means the 'what' and 'where' in the scene. Monomodal perceptual driven video adaptations mainly regard:

- **Temporal selectivity -** The adaptation represents with better fidelity/quality a specific temporal period of the content, e.g., a specific event, with the target to improve the quality in the temporal periods with more relevant information. Also if there is a duration constraint, the adaptation may remove the parts of the video less relevant in terms of informative content.
- **Spatial selectivity -** The adaptation represents with better fidelity/quality, better contrast, etc., a specific spatial area of the content, e.g., a region of interest in a picture, with the target to improve the quality in the spatial areas with more relevant information for the cognitive experience.
- **Text selectivity – T**he adaptation represents with a more adequate color or size the text in the content to improve its readability.
- **Scene composition selectivity -** The object-based scene is represented with a reduced number of objects, e.g. only high priority objects, with the target to maximize the cognitive experience understood as a combination of sensations and perceptions, e.g., it is more informative to have less objects with a minimum quality than more objects with unacceptable quality.

Transmodal perceptual driven video adaptations are mainly driven by the type of resource constraints, notably:

- **Device modalities –** If a certain modality cannot be presented at the available user device then a more informative experience will be to present a transmoded version of the content using a modality or modalities that can be presented at the device in question.
- **Bitrate –** If a low bitrate is available then a set of images, a speech, or a text description may provide a more informative experience than the transcoded video.
- **Screen size –** If an image cannot fit with an acceptable resolution into a screen, it may be transformed into an attention-based created video which allows an easier and more efficient cognitive experience of the content (e.g., compared with scrolling the image left-right and up-down).

The subjective assessment of perceptual quality may be performed in a rather global way, e.g. asking the user questions like "Do you think the two versions (adapted and non-adapted) have the same informative value?" or using task-based evaluation where content specific questions (associated to relevant tasks) are developed for the non-adapted version, e.g. "How many cars do you see in the street?".

The objective assessment of perceptual quality is even more difficult than the sensorial assessment and has typically to consider a temporal and a spatial dimension. For example, it is clear that the informative value of a video sequence may be reduced if some shots or frames are removed but this reduction strongly depends on the parts removed. In the same way, the informative value of a video sequence may be reduced if some objects are removed from a scene or if some objects are adapted to a size or a quality which limits the cognitive experience. In [3], a shot utility function is defined as the product of the shot duration and complexity (this may be associated to motion activity, spatial variation, texture type, etc.) which gives a first prediction of the perceptual value of a shot. This metric has limitations in terms of perceptual evaluation, for example because it does not consider the informative redundancy between shots. For an object-based scene, this metric may be evaluated object by object, eventually weighting more the evaluation for specific objects, e.g., faces or text, if a face or text detector is also included in the process. The weight for each object may depend on its semantic relevance, position in the scene, size, etc.

For a single object-based shot with N objects, a perceptual quality metric may be

$$PQ_{shot} = \sum_i^N w_i\,(relevance, position, size)\,.\,f_{1i}(duration, n^o\,keyframes)\,.\,f_{2i}\,(motion, texture)$$

with the perceptual quality for the sequence coming

$$PQ_{sequence} = \frac{1}{f\,(duration_1 ... duration_N\,)} \sum_j PQ_{shot\ j}\ w_j\ (duration_{\ j})$$

where the second factor under the sum weights the influence of the different duration of the various shots, in a linear or non-linear way, and f (duration$_1$, … duration$_N$) normalizes the overall perceptual quality. A more sophisticated metric could weight the perceptual quality for each shot depending on its perceptual relevance; for example, in certain contexts a more active shot is typically more relevant and thus its perceptual quality should be higher.

3.3 Adaptation and quality at the emotional layer

Adaptation at the emotional level may involve either transcoding or transmoding always with the target to maximize the emotional experience. This type of adaptation is the only with the novel characteristic that may present the user more information than was originally available (which is a rather unusual feature in today's available solutions). An example is image to video (generated from the image through attention based mechanisms) plus music (selected after affective analysis of the image) transmoding where inclusion of music (not part of the original content) in the experience may increase the quality of the experience.

While in a first approach the adapted content with additional modalities may be simply seen as a new, different piece of content this may not be the case because i) the additional modalities do not intend to contribute to the informative component of the content; ii) the additional modalities are not predetermined, neither in type, nor in specific piece, being dependent on the adaptation mechanism at hand, iii) the additional modalities do not have even to be transmitted by the sender and may be locally added by the presentation application.

Monomodal emotionally driven video adaptations may regard:

- **Colour temperature –** The adaptation transforms the pixels to provide a warmer or colder feeling to the user.
- **Temporal selectivity** - The adaptation represents with better fidelity/quality an emotionally more relevant temporal period of the content, e.g., a specific event, with the target to increase the intensity of the emotional experience. Also if a duration constraint exists, the adaptation may remove the parts of the video which are less relevant in terms of emotional content.
- **Spatial selectivity –** Same as above for the spatial dimension.
- **Scene composition selectivity –** Same as above for the scene composition.

Transmodal emotionally driven video adaptations may regard:

- **Change of modality(ies) –** The adaptation presents to the user the same information using a different modality which is able to create a more emotional experience, e.g., transforming an image into a video following some relevant criteria.
- **Addition of modalities –** The adaptation presents to the user the original modalities part or all of which transcoded or transmoded based on perceptual motivations but also one or more additional modalities included for the exclusive reason to increase the user satisfaction by means of a more intense emotional experience; an example is again the display of images from a database adding adequate music after affective analysis of the images.

It is clear that emotion-based adaptations are those where creativity may play a bigger role and also where the user appears in his/her full capacity as a human being. Also this is the type of adaptation where culture, as well as all types of sociological and psychological issues may impact more on the precise adaptation to be performed. Finally, it has to be recognized that more acute intellectual property rights issues may arise, notably for the case where new modalities are added.

The best way to measure emotional quality is again subjective and this may be performed in a more global or specific way, e.g. “Do you think the two versions (adapted and non-adapted) have the same emotional impact?” or “How high is your feeling of happiness?” Regarding the objective evaluation of emotional quality, and opposite to the previous cases, it is more centered on the user than on the content and thus it is difficult to measure it directly using the content. This means this type of objective assessment has to be based on user reactions measured in a more or less invasive way, e.g., cardiac beat, skin conductance, temperature, muscular and cerebral activity, brain magnetic resonance imaging, facial emotions. Functional magnetic resonance imaging (fMRI) is used to visualize brain function and some have succeeded in using this technique to study emotional processes. One of the theories to describe emotions is Plutchik's theory of emotion. Plutchik's model is based on an emotion wheel which shows eight basic emotions made up of four pairs of opposites: joy and sadness, acceptance and disgust, fear and anger, and surprise and anticipation. Emotional strength which may be related to emotional quality can be objectively measured as the distance from the emotion wheel origin to any point in the Plutchik's emotion wheel. This emotional strength may be measured, for example, by means of automatic facial expression analysis but this is where research on content adaptation starts touching the limits of knowledge in other scientific areas … this is where we will work in the next years.

3.4 Metrics for quality of experience

Although the major objective of this paper is not to propose quality of experience metrics, it is worthwhile to throw here the first stone in terms of triple model quality metrics.

Since a low quality experience in one quality dimension should make the whole quality of the experience low, two simple metrics with different characteristics that can be used to measure the overall quality of experience are:

$$QoE = SQ^{ws} \cdot PQ^{wp} \cdot EQ^{we}$$

or $QoE = w_s \ SQ \,.\ w_p \ PQ \,.\ w_e \ EQ$

where w_s, w_p and w_e express the weights of the sensorial, perceptual and emotional dimensions in the overall quality of experience since, depending on the type of service, one or more of these quality dimensions may have more weight. For example, the emotional dimension is certainly more relevant for entertainment services than for video surveillance. Also, the quality of experience metric may mix (normalized) subjective and objective evaluation scores for the three quality evaluation dimensions depending on the relevant evaluation limitations. While there are already good enough objective assessment solutions for the sensorial layer, the same does not happen for the other 2 layers which may still have to rely mostly on subjective assessment. An interesting alternative to the metrics above are similar metrics with explicit emotion and non-emotion-based quality components such as:

$QoE = SQ^{w_{s1}} \cdot PQ^{w_{p1}} \quad + \quad SQ^{w_{s2}} \cdot PQ^{w_{p2}} \cdot EQ^{w_{e2}}$ or

$$QoE = w_{s_1} \ SQ \,.\, w_{p_1} \ PQ \quad + \quad w_{s_2} \ SQ \,.\, w_{p_2} \ PQ \,.\ w_{e_2} \ EQ$$

This type of metrics may be more adequate for services for which there are clearly two quality components which depend differently on emotional factors and also have a different relation between the three quality dimensions.

4. REFERENCES

[1] F.Pereira, I.Burnett, “Universal multimedia experiences for tomorrow”, IEEE Signal Proc. Mag., Special Issue on Universal Multimedia Access, vol.20, nº 2, pp. 63-73, March 2003.

[2] S. Coren, L. M.Ward, J. T. Enns, Sensation and perception, Harcourt Brace College Editors, 1994.

[3] H. Sundaram, L. Xie, S. Chang, “A utility framework for the automatic generation of audio-visual skims”, ACM Multimedia, Juan Les Pins, France, December 2002.