A SUBJECTIVE STUDY OF THE INFLUENCE OF COLOR INFORMATION ON VISUAL QUALITY ASSESSMENT OF HIGH RESOLUTION PICTURES

Francesca De Simone[a], Frederic Dufaux[a], Touradj Ebrahimi[a], Cristina Delogu[b], Vittorio Baroncini[b]

[a]Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
[b]Fondazione Ugo Bordoni (FUB), Via Baldassarre Castiglione 59, 00142 Rome, Italy

ABSTRACT

This paper presents the design and the results of a psycho-visual experiment which aims at understanding how the color information affects the perceived quality of a high resolution still picture. The results of this experiment help to shed light into the importance of color for human observers and could be used to improve the performance of objective quality metrics.

1. INTRODUCTION

The compression efficiency of an image coding algorithm expresses its ability to maximize the visual quality of a compressed image, versus the number of bits used to represent it, for a range of compression ratios. As human subjects often act as the end users of the digital content, subjective tests can be performed, where a significant sample of human subjects is asked to rate the quality of the processed material.

Since these tests are time consuming and expensive, usually objective metrics are used in order to assess the quality of the compressed images. Such metrics are called Full-Reference (FR) quality metrics, because they assume as input both the original image (i.e. reference) and its compressed version. A substantial effort has been recently deployed by the research community to design objective visual quality metrics which achieve a good correlation with the subjective quality evaluation. Nevertheless, most of the well-known and widely used FR quality metrics take into account only the luminance channel of the picture under analysis [1], i.e. they are "luminance only metrics" or "single-channel metrics". Examples of color quality metrics include the fidelity metric developed by Chou *et al.* [2], designed to measure the perceivable distortion for each color pixel in the quasi-uniform CIELab color space, the color image quality metric by Le Callet *et al.* [3], which relies on a psycho-visual representation stage of the reference and test images, including the color information of the data, and on an error pooling stage based on error density and error structure, and the psycho-visual color image quality metric designed by Charrier *et al.* [4], which takes into account the human color contrast sensitivity when computing the distortion measure.

Major drawback of these "color metrics" or "multi-channel metrics" is that only very few verification results are presented. These results do not allow to conclude that a significant improvement is achieved in terms of correlation with the subjective quality perception, with respect to much simpler luminance only metrics. Furthermore, these algorithms are quite complex and their implementations are not publicly available.

Hence, the quality performance evaluation and optimization of full color image algorithms are usually done by means of methods which are applied on the luminance component only. Assuming that the color information significantly influences the human visual quality perception, it can be easily concluded that this approach is limiting *a priori* the correlation that can be met with such single-channel metrics compared to the subjective judgment. Unfortunately, at the best of author's knowledge, studies analyzing the influence of the color information on the perceived visual quality of digital images are not available in literature.

This paper aims at providing a first contribution in this direction, presenting a study of the influence of color information upon the subjective perception of quality of high resolution still pictures. A methodology is designed to perform a psycho-visual experiment with human subjects so as to understand *"how the color information affects the overall assessment of the distorted image with respect to the perceived quality on the luminance only version of the same distorted image"*. The information gathered through this experiment helps in understanding and quantifying the margin of improvement which could be achieved by including the color information in the objective quality evaluation models. This knowledge could be used in order to develop multi-channel objective quality metrics which better predict the human quality judgment.

The test methodology and the selection of the test material are detailed in Section 2 and Section 3, respectively. The processing applied to the subjective data resulting from the experiment is described in Section 4. The results of the study are presented in Section 5. Finally, conclusions are drawn in Section 6.

2. EXPERT VIEWING TEST METHOD

The goal of this experiment is to provide a proof of the different ability of the human subjects to detect visual impairments when assessing a full color image or only its luminance version. The lack of standards defining how to test mono-channel (luminance) vs. multichannel (color) pictures and the high resolution of the considered images make the selection of the test method rather difficult. To overcome this issue, it was decided to refer to the experience of ITU-R in the area of subjective evaluation of television images and in particular to the DSCQS (Double Stimulus Continuous Quality Scale) method, described in [5]. The choice of the DSCQS method is suggested by the suitability of this method for the evaluation of high quality moving images in presence of an unimpaired reference sample.

The test method used here is derived from the DSCQS test method, with minor adaptations dealing with:

- the evaluation of fixed images;
- the use of an interactive method of data collection.

The evaluation of fixed images is different from the evaluation of moving images mainly for the way the eyes explore the visual area. When assessing a moving image, the fovea tracks the area of major interest in the scene. On the other side, when evaluating fixed images, the eyes explore in a more extensive way the whole visual area, attempting to get the sensation of maximum resolution for any portion of the observed image.

The above considerations lead to selection of a test methodology in which each human subject is left free to examine a test picture and its unimpaired version as long as she/he gets the sensation of having completed the quality investigation task. For this reason, a dedicated GUI was created to allow the human subjects to assess pairs of pictures, with a reference unimpaired picture on one side of the screen, and a degraded version on the other.

The experience in the assessment of high quality video images during the activities of ITU-R SG6 Task Group 6/9 "Digital Cinema" suggested the usage of experts other than naïve viewers. This previous experience allowed to conclude that the assessment by a group of experts is highly reliable and able to provide results as reliable and stable as those obtainable performing a standard subjective test. This conclusion leads to the approval of the recommendation dedicated to the expert viewing of Large Scale Digital Imagery (LSDI) [6].

Thus, an "expert viewing" test is performed. This choice allows speeding up the test. Six experts screened for visual acuity (Snellen chart), and color blindness (Ishihara tables), participated in the tests during two successive days.

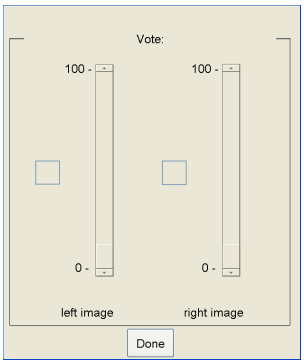

Figure 1. Voting window

2.1 Test timing and displays

As the test session starts, a window is displayed allowing the subject to fill in date, time and her/his name. After this preliminary step, the test begins. The first pair of images is displayed. When the subjects is ready to vote, she/he clicks into the active area of the screen and the voting window is shown (see Figure 1). When the subject has moved the slider to the desired position (a small window shows the numerical value corresponding to the slider position) she/he confirms the judgment clicking the "Done" button. The scale available to express the votes ranges from 0 (Very bad quality) to 100 (Excellent quality) according to the DSCQS method.

The test material is presented in a pseudo-random order so that test pairs related to different original contents are always alternated, i.e. test pairs related to the same original content are never presented on two successive occasions with the same or different levels of impairment. Before each test session, written and oral instructions are provided to the subjects to explain their task. Additionally, a training session is performed to let the subject familiarize with the interface and to explain her/him how to use the rating scale. The contents shown in the training session are not used in the test session and the data gathered during the training are not included in the final test results.

The test is performed by considering a first session called "color data session", where color images are displayed (i.e. reference image and distorted version in the same screen), and a second session, called "luminance data session", where just the luminance component of each color image of the previously used dataset is displayed (i.e. luminance component of the reference shown together with the luminance component of the distorted image).

The duration of each session is highly depending on the time each subjects dedicates to the viewing of each image. No limit is given in this sense, leaving each subject free to spend as long as she/he retains to be necessary in order to perform a correct evaluation. The color and luminance sessions are performed by each subject in two different days. This is to avoid the influence of a short term memory.

2.2. Test room set up

The experiment was conducted at the Fondazione Ugo Bordoni test laboratory. The highly professional laboratory set-up is designed in full agreement with the relevant recommendation issued by ITU-R for the subjective evaluation of fixed and moving images [5] [6]. The laboratory set-up is intended to assure the reproducibility of the subjective test activity by avoiding the involuntary influence of any controllable external factors. To allow these results the test was conducted in a room cleared from any visual and audible pollution.

A high performance PC was used. The video board was connected to two identical high resolution monitors (Samsung 226 CW). Each monitor was verified and calibrated using a color calibration device (Eye-one Display 2). The video board monitor handling function was set up to extend the desktop area to both monitors. In this way, a desk top of 3200 x 1200 pixels was obtained. A GUI presented the original image on one screen and the corrupted image on the other screen, as two adjacent windows of equal size.

3. TEST MATERIAL SELECTION

The choice of the input data to be used in the psycho-visual experiment is clearly related to the design of the methodology and it is a critical point affecting the significance of the collected subjective data. As a starting step, we restricted our field of investigation by considering a dataset of pictures including only natural images and the distortions introduced by two compression algorithms specified hereafter. The selected pictures are 8 bits per channel, i.e. 24 bit-per-pixel (bpp), high resolution pictures chosen from the database established by Microsoft [7]. Since this database includes very high resolution pictures (up to 4288x2848 pixels), a central selected crop of each picture, of 2560x1600 pixels, has been considered as the original image. In particular, twenty-three natural pictures have been considered and cropped. Thumbnails of these crops are shown in Figure 3. The coding is applied to these selected crops.

To allow the previously described presentation of the test material, where the reference image and the test image are shown on the same screen at the same time, the subjective data are collected by presenting only one part of each test picture corresponding to half of the screen resolution, i.e. two 1600x1200 images. As detailed hereafter, a preliminary informal test has shown that the results obtained by displaying only one part of the picture are correctly approximating the results obtained if the entire picture would be shown. The procedure applied to select the original contents and the compression rates of the test pictures is detailed in the following subsections.

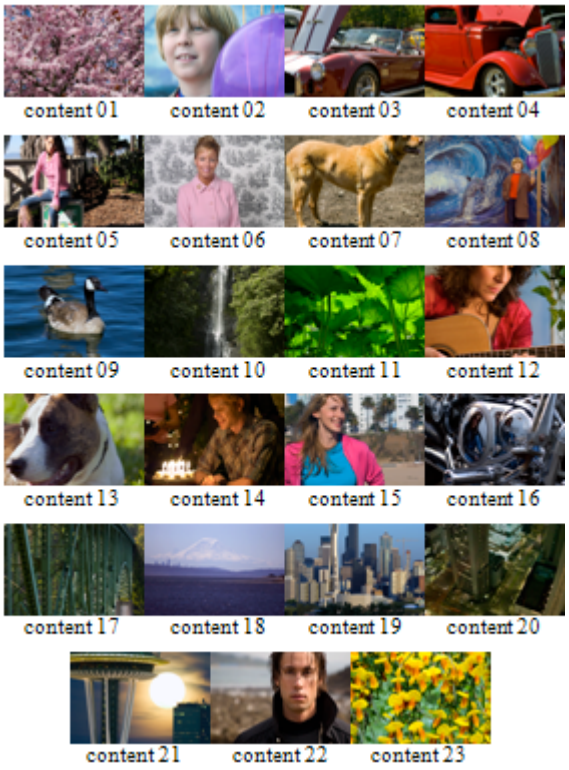

Figure 2. Thumbnails of crops of twenty-three images from [7], considered in order to select significant original test contents.

3.1. Selection of the original contents

The need to reduce the duration of each test session implies a strong limitation in the number of original contents which can be included in the test material. In order to select significant test images, the twenty-three pictures originally considered have been classified by estimating “how difficult each image is for coding”. Then, images which show clearly different coding complexity levels have been chosen as test images.

The level of coding difficulty has been estimated by analyzing the slope of the Rate-Distortion (RD) curves produced for the luminance and the two chrominance components of each original image, using the two compression algorithms specified in the next subsection. The Mean Structural SIMilarity (MSSIM) index [8] and the Peak Signal to Noise Ratio (PSNR) have been considered as distortion measures. By following this procedure, four different classes of pictures have been identified, mainly based on the slope of the RD curves for the luminance channel of the pictures. This process has also been validated by an expert view, visually checking the quality variation of the compressed pictures. Additionally, a Spatial Information (SI) index has been computed for each component of the image as: SI=std[Sobel(ImComponent)], that is the standard

deviation (std) computed over the pixels values of the Sobel-filtered component of the image [9]. The SI index provides information upon the spatial complexity of each component. Apart from some outlier cases, the SI on the luminance component shows a good approximation of the coding complexity of the image.

Selected image	Coding complexity level	SI index (Y component)
Image 09	1 (simple)	3.2159
Image 12	2	4.7696
Image 23	3	6.0753
Image 17	4 (difficult)	8.2654

Table I. Coding complexity levels and SI indexes of the four images selected as original test contents.

Performing this semi-automatic content classification, four images have been identified as representatives of four different levels of coding complexity and selected to be used as original test contents, namely images 09, 12, 17 and 23 shown in Figure 3. The coding complexity level and the SI indexes of these four images are shown in Table I.

3.2. Selection of the compressed pictures

The compressed pictures used in the test were produced by coding the original contents in the Y'CbCr color space [10] without spatial sub-sampling of the chrominance components (i.e. 4:4:4 coding) using i) JPEG with conventional visually optimized quantization matrix [11] and ii) JPEG 2000 with frequency weighting of quantization steps [12]. These two coding algorithms have been selected because they produce different kinds of artifacts: while blocking and false contouring artifacts are mainly present in JPEG compressed pictures, JPEG 2000 compressed pictures show blurring and ringing artifacts.

First, for each coding algorithm, the same number of compressed samples is selected. In particular, in order to have a dataset of reasonable dimension, a preliminary selection of the test material led to selection of five samples corresponding to five different levels of quality of the compressed content, for each coding algorithm. The selection of the test samples is performed only on the color data and in an environment having exactly the same characteristics as the test environment detailed in subsection 2.1.

An expert viewer performs the "visual area identification" for each content and coding condition: the set of compressed pictures is visually checked in order to identify the quality saturation values in terms of minimum and maximum bpp values before which, and after which, respectively, the quality does not change anymore. This range of bpp values delimits the so-called "visual area" of the compressed images set.

The following procedure is then applied to select the significant samples in the visual area:

1) Pairs of pictures are shown on the screen, having on one side the Reference Picture (RP) and on the other side the Comparison Picture (CP).
2) The expert viewer is asked to indicate whether she/he notices a difference in terms of overall picture quality between the two samples.
3) When a difference is detected, the current CP is stored as representative of one quality level, and becomes the reference while a new CP is loaded and shown. If no difference is detected the reference remains the same while a new comparison picture is shown.

The first reference picture shown is the original uncompressed picture. The comparison pictures are selected from the set of compressed pictures, starting from the lowest compression ratio, i.e. highest bit per pixel value, and sequentially up to the highest significant compression ratio, i.e. smallest bit per pixel value. The different original contents and coding conditions are analyzed in separate sessions.

At the end of this procedure, for each content and coding condition, a certain number of pictures, i.e. bpp values, have been selected as representative of a corresponding number of quality levels. If more than five different quality levels are detected, for one content or coding condition, the procedure is repeated a second time. This time the pictures which are compared are only those which have been chosen as representative of quality levels in the previous selection.

3.3 Final data set

To further reduce the time and complexity of the test, out of the ten compressed pictures selected for each content as described above (i.e. five compressed pictures produced using JPEG coding and five compressed pictures produced using JPEG 2000 coding), three samples have been discarded. In this way, only test pictures having quality levels which are reasonably distinguishable from each other are considered. In particular, as shown in Section 5, three of the compressed test images have been produced using JPEG, while the remaining four have been produced using JPEG 2000.

The dataset for the "color test session" is thus composed of four original test images and seven compressed versions for each of them, i.e. 28 different test pairs. The dataset for the "luminance data session" is simply obtained by applying the RGB to Y'CbCr color transform [10] to the 28 test pictures and considering only the luminance component of each picture.

3.3.1. Focus of attention in high resolution images

As mentioned at the beginning of the section, when two versions of the same content have to be shown at the same time on the monitor, only one part of each picture can be shown in order to fit the native resolution data in the screen. Due to the homogeneity of the characteristics of the content,

for image 09, 17 and 23, showing either the right part or the left part only, rather than the entire picture, does not have any effect on the quality judgment.

When considering image 12 instead, the content features are not homogeneous in the two sides of the image. In this case, a slight overestimation of the quality of the full resolution picture occurs if only the less difficult side of the image is shown. In fact, the analysis of high resolution images requires the subject to move her/his head to scan the entire content. This scanning process is uncomfortable for the subject. For this reason, the part of the image which influences the most the quality judgment of the entire image is the area which attracts the most the visual attention of the subject. This influence is much stronger than in the quality assessment of a standard resolution image, since in that case, apart from a focus on the most attractive area of the image at a first glance, all the content is easily scanned and analyzed with accuracy.

The attractiveness of a visual area can be identified by applying a visual Focus of Attention (FoA) model like the one designed by Itti *et al.* [13]. Considering content 12, showing only the left side of the image allows obtaining quality judgments which are correct approximations of quality judgments of the entire image.

4. SUBJECTIVE DATA PROCESSING

For each original content, the Differential Mean Opinion Score (DMOS) was computed for each test condition. DMOS is the result of the separate evaluations of the original and of the coded samples of the same picture. To obtain the DMOS value, the MOS value assigned to the coded picture is subtracted from the MOS value assigned to the original picture. This leads to an inversion of the quality index meaning between the MOS and the resulting DMOS values. In other words, as MOS values go higher, the visual quality is judged higher, whilst for the DMOS higher values of visual quality correspond to lower scores.

Two sets of subjective results are obtained: one set of DMOSs values is referred to the quality evaluation of luminance only stimulus. We will refer to it as DMOSluma. The other set is related to the quality assessment of the color stimulus, referred to as DMOScolor. The results are grouped for each test image, as shown in the following section. Due to the fact that the test was done using only six experts, it was not possible to perform an analysis of variance and therefore no values for standard deviation and confidence intervals are available.

5. RESULTS

The graphs in Figures 3-6 show, for each original content (image 09, 12, 17 and 23 respectively), the DMOSluma and DMOScolor values of the seven compressed test samples, indicated as J_01, J_02, J_03, J2_01, J2_02, J2_03, J2_04. The samples J_01 to J_03 are JPEG compressed images, with descending compression ratios. The samples J2_01 to J2_04 are JPEG 2000 compressed images, with descending compression ratios. The values on the abscissa are ordered to present ascending DMOSluma values in the graphs.

The data clearly shows that for the majority of the cases, the full color images are judged to be lower in quality than the equivalent luminance only samples, independently from the original content under analysis. Usually no difference is present when the quality of the compressed image is too high or too low.

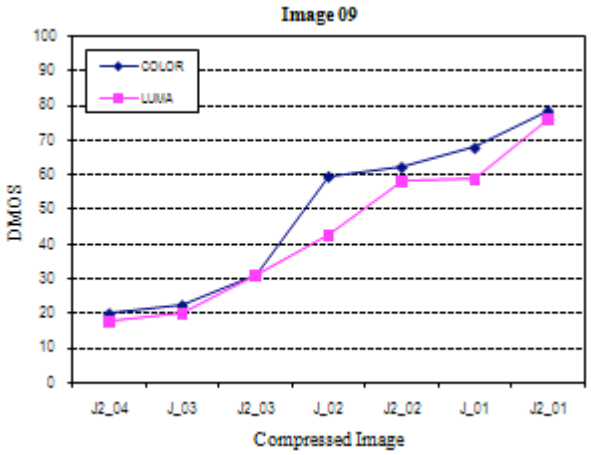

Figure 3. Results for image 09

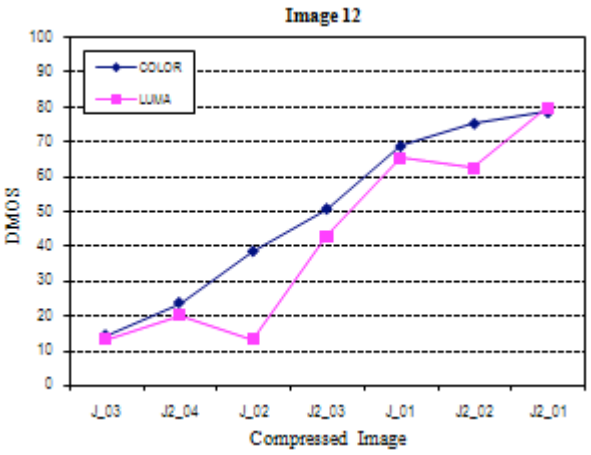

Figure 4. Result for image 12

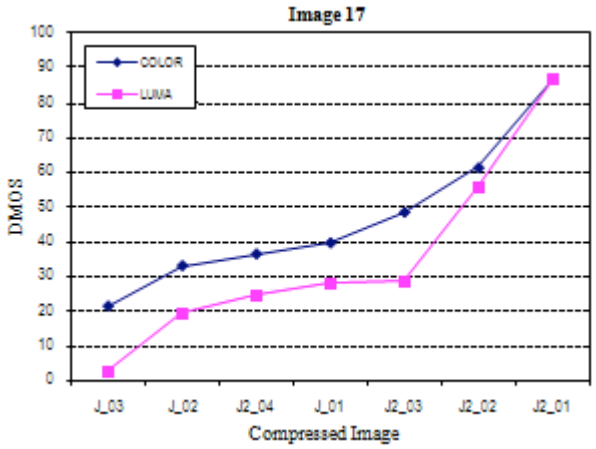

Figure 5. Result for image 17

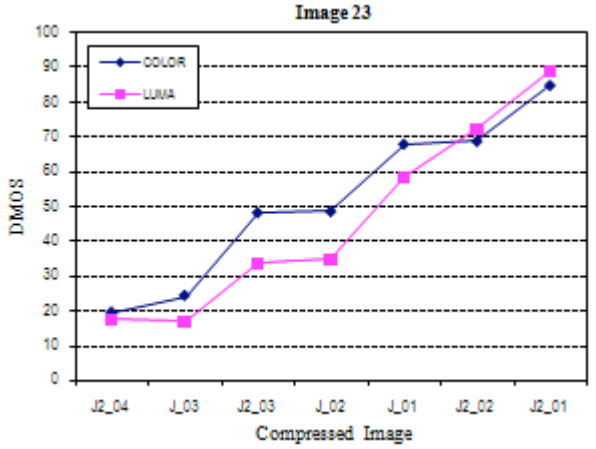

Figure 6. Result for image 23

6. CONCLUSIONS

This paper proposes an investigation upon the influence of the color information on the visual quality level perceived by a human subject when considering high resolution natural images. In particular, the results of this investigation show how a relevant difference in the visual quality of the data is perceived by subjects if only the luminance channel of the image is shown rather than the full color stimulus. This evidence should encourage the design of objective visual quality metrics which include the information of the chrominance channels, in order to achieve a better correlation with the end user quality judgment compared to the objective metric applied on the luminance channel only. Future works will focus on extending the panel of subjects of the proposed test methodology, including naïve viewers. Additionally, a deeper characterization of the original contents will be performed, taking into account not only the spatial features but also the color and contrast features of the images. Finally, an analysis of artifacts masking effects associated with the obtained results will be performed.

7. ACKNOWLEDGEMENTS

The work presented here was partially supported by the European Network of Excellence VISNET II (IST Contract 1-038398), the European Network of Excellence PetaMedia (FP7/2007-2011), and the Swiss National Foundation for Scientific Research in the framework of NCCR Interactive Multimodal Information Management (IM2). A special thank also goes to the six experts who participated with high dedication to the subjective test reported in this paper.

8. REFERENCES

[1] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," IEEE Trans. Image Process., vol. 15, pp. 3440 – 3451, Nov. 2006.

[2] C. Chou, K. Liu, "A Fidelity Metric for Assessing Visual Quality of Color Images", in Proc. of 16th IEEE Int. Conf. on Comp. Comm. and Net. (ICCCN), pp. 1154-1159, Aug. 2007.

[3] P. Le Callet, D. Barba, "A robust quality metric for color image quality assessment", in Proc. of Int. Conf. on Image Process. (ICIP), vol.1, pp. 437-40, Sept. 2003.

[4] C. Charrier, T. Eude, "A psychovisual color image quality metric integrating both intra and inter channel masking effect", in Proc. of 12th Eur. Sig. Proc. Conf. (EUSIPCO), Sept. 2004.

[5] ITU-R Rec. BT.500-11, "Methodology for the subjective assessment of the quality of television pictures".

[6] ITU-R Rec. BT.1663 "Expert viewing methods to assess the quality of systems for the digital display of LSDI in theatres".

[7] Dataset available for JPEG members at http://www.jpeg.org.

[8] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol. 13, no. 4, pp. 600-612, Apr. 2004.

[9] ITU-T Rec. P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications", Sept. 1999.

[10] ITU-R Rec. BT.601-6, "Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratio".

[11] http://www.ijg.org/

[12] http://www.kakadusoftware.com.

[13] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems", J. Electr. Im, Vol. 10, pp. 161-169, 2001.