View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Infoscience - École polytechnique fédérale de Lausanne

Multimed Tools Appl
DOI 10.1007/s11042-012-1011-6

Paired comparison-based subjective quality assessment of stereoscopic images

Jong-Seok Lee · Lutz Goldmann · Touradj Ebrahimi

© Springer Science+Business Media, LLC 2012

Abstract As 3D image and video content has gained significant popularity, subjective 3D quality assessment has become an important issue for the creation, processing, and distribution of high quality 3D content. Reliable subjective quality assessment of 3D content is often difficult due to the subjects' limited 3D experience, the interaction of multiple quality factors, minor quality differences between stimuli, etc. Among subjective evaluation methodologies, paired comparison has the advantage of improved simplicity and reliability, which can be useful to tackle the aforementioned difficulties. In this paper, we propose a new method to analyze the results of paired comparison-based subjective tests. We assume that ties convey information about the significance of quality score differences between two stimuli. Then, a maximum likelihood estimation is performed to obtain confidence intervals providing intuitive measures of significance of the quality differences. We describe the complete test procedure using the proposed method, from subjective experiment design to outlier detection and score analysis for 3D image quality assessment. Especially, we design the test procedure in a way that quality comparison across different contents is enabled while the number of pair-wise comparisons is minimized. Experimental results on a stereoscopic image database with varying camera distances demonstrate the usefulness of the proposed method and enhanced

J.-S. Lee (✉)
School of Integrated Technology, Yonsei University, 406-840 Incheon, Korea
e-mail: jong-seok.lee@yonsei.ac.kr

L. Goldmann · T. Ebrahimi
Multimedia Signal Processing Group (MMSPG), Swiss Federal Institute of Technology in Lausanne (EPFL), 1015 Lausanne, Switzerland

L. Goldmann
e-mail: lutz.goldmann@epfl.ch

T. Ebrahimi
e-mail: touradj.ebrahimi@epfl.ch

Published online: 11 February 2012

quality discriminability of paired comparison in comparison to the conventional single stimulus methodology.

Keywords Stereoscopic image · Subjective quality · Paired comparison · Quality of experience (QoE)

1 Introduction

Nowadays, three-dimensional television (3DTV) is receiving a great deal of attention as a new multimedia experience for consumers, and a significant amount of research has been conducted for 3D content creation, processing, distribution, and restitution [9, 16, 21]. In order to reach the goal of 3D content delivery with satisfactory quality of experience (QoE), reliable subjective and objective quality evaluation is essential. The issue of assessing perceived quality is particularly important in 3D content because, unlike 2D quality, degraded 3D quality may cause not only dissatisfaction but also serious discomfort [14, 17].

In comparison to conventional 2D content, knowledge on perceived quality of 3D content is relatively insufficient. Thus, extensive subjective quality evaluation studies are crucial for understanding the way that human subjects perceive 3D content, which will be eventually useful for the optimization of 3D processing techniques and the development of objective 3D quality metrics [10, 11]. Existing subjective quality assessment methods can be classified into psycho-perceptual and user-centered approaches. The former quantitatively examines the relation between physical stimuli and sensory experience through controlled experiments, usually in laboratory environments. Typically, subjects in such experiments are asked to provide quality scores for presented stimuli, and the collectied scores are analyzed to obtain quantitative quality measures. The latter aims at evaluating the quality from a user's perspective by considering the potential use (such as users, system characteristics and context of use) [13], where unconventional procedures such as field experiments, questionnaires and interviews are often involved. While the user-centered approach may allow more in-depth investigation of quality perception in particular scenarios, the psycho-perceptual approach is still preferable due to the easy quantitative analysis and reproducibility of the results. This paper focuses on the former approach.

When one conducts subjective quality assessment, it is necessary to carefully design a test procedure and subsequent methodology suitable to the chosen goal of the study in order to obtain reliable quality ratings from subjects without unwanted external factors and bias in human perception. International standards provide guidelines of test methodologies for subjective tests [18, 23]. While single stimulus and double stimulus methodologies are widely used, paired comparison has its own advantage of simplicity, i.e. only the preference of a subject between each pair of stimuli is asked for instead of a score of an individual stimulus in a discrete or continuous scale.

The simplicity of the paired comparison methodology promotes itself as an effective alternative for subjective evaluation of 3D content. 3D quality evaluation for naive subjects is not as easy as 2D quality evaluation, because they do not have sufficient experience of 3D content in daily life and thus quantitative judgment of

3D quality becomes very difficult. This sometimes leads to unreliable results with large variances across subjects in quality scores. In such cases, paired comparison significantly reduces the complexity of the subjects' task and, consequently, enhances reliability of the results.

A challenge in employing paired comparison is to obtain absolute quality scores from the comparison results. The results of a paired comparison experiment with multiple subjects appear as winning frequencies of each stimulus against each of the other stimuli. In order to facilitate quality assessment of the stimuli, they need to be translated into quality scores that are equivalent to mean opinion scores (MOS) of single stimulus methodologies or differential mean opinion scores (DMOS) of double stimulus methodologies. In addition, it is desirable to obtain confidence information of the quality scores for further analysis of the results. In single or double stimulus methodologies, a confidence interval for each MOS or DMOS is commonly computed for a chosen significance level.

In this paper, we propose a new analysis method for paired comparison, which estimates quality scores from winning frequencies and, more importantly, extracts significance information of the scores to facilitate intuitive examination of significant quality differences between scores. The method is applied to subjective quality assessment of stereoscopic images, and in particular, for varying distances of stereoscopic cameras used during the acquisition of the images. The detailed description of the subjective experiment shows how the paired comparison methodology can be effectively used for quality evaluation of stereoscopic images and how the results are analysed using the proposed method. The results of the experiments demonstrate consistency between the results of the paired comparison and conventional single stimulus methodologies and decreased ambiguity between similar quality levels in paired comparison.

The remainder of this paper is organized as follows. The following section presents our proposed framework for paired comparison-based subjective quality assessment. In Section 3, we present our study where the proposed framework is applied to quality assessment of stereoscopic images acquired with a varying acquisition parameter. Finally, conclusions are drawn in Section 4.

2 Proposed subjective quality assessment framework

2.1 Overview

In our framework, the task of a subjective quality assessment experiment is to obtain quality scores of stimuli that belong to different "groups" and have been generated for varying "conditions". For example, when the perceived quality of images contaminated by noise is to be investigated using diverse image contents, the group and condition correspond to the content and type/amount of noise, respectively.

In conventional single or double stimulus methodologies, the test for a subject includes evaluation of all or part of a set of stimuli, which results in quality scores given by the subject for the considered stimuli. However, the number of pairwise comparisons increases exponentially with respect to the number of considered stimuli and, thus, it is necessary to devise a systematic way to reduce the number

of comparisons for each subject but to make it still possible to estimate globally comparable quality scores of all the stimuli. Therefore, a test based on the proposed framework consists of two phases, namely, intra-group comparison and inter-group comparison. In the *intra-group comparison*, a pair-wise comparison is conducted for all possible pairs in each group independently. The results of the comparisons appear as *winning frequencies*, i.e. frequencies that a stimulus wins the other stimuli for multiple subjects. They are converted into relative quality scores of the stimuli in each group by the process described in Section 2.4. Then, the *inter-group comparison* is performed for the stimuli of some selected conditions across the groups. The winning frequencies obtained in this phase are converted into relative scores in a similar way to that used in the intra-group comparison phase. These results are used to globally align the quality scores of all the involved stimuli, which will be explained in Section 2.5.

Note that the inter-group comparison phase is rather optional and may be skipped in some cases. For example, the task of the study presented in [15] was to examine which video scalability option among spatial, temporal and quality dimensions is important for maximizing perceived quality for each fixed bit rate constraint. In this case, comparing quality across different bit rate constraints was not of interest, so that the inter-group comparison, i.e. comparison across bit rate constraint, was not necessary. However, in the particular task considered in this paper, a group corresponds to a scene and the inter-group comparison plays an important role because we are interested in investigating the effect of content features to the perceived quality.

2.2 Test procedure

In a paired comparison test, a subject is asked to provide an index of the relation between two stimuli presented. The judgment can be done in various ways. A score value on a continuous scale can be given to describe the relation, e.g. $[-100, 100]$, where -100 (or 100) indicates that the quality of stimulus A (or B) is much lower than that of stimulus B (or A). Or, a categorical judgment can be used by selecting one of predefined semantic categories, e.g. the simplest form (stimulus A is 'better' or 'worse' than stimulus B) and a more subdivided form (stimulus A is 'much better', 'better', 'slightly better', 'same', 'slightly worse', 'worse', or 'much worse' than stimulus B). In order to keep the task simple, only three judgment options are considered in our case, namely, 'A is better', 'B is better', and 'same'. In comparison to single or double stimulus methodologies, the employed paired comparison methodology improves the simplicity of the subjects' task significantly due to the simplified rating scale. Moreover, the simplicity is improved further when simultaneous viewing of two stimuli is enabled, as in our study presented in Section 3.

Inclusion of a tie in the rating scale prevents biased results due to random choices between two stimuli when their quality difference is hardly noticeable. Such a bias may be also avoided by considering a large number of subjects and/or repeating the same pairs several times in a test, which is inefficient and even inapplicable in most cases. Moreover, a tie conveys information related to uncertainty of the quality difference of the two stimuli, and thus can be used to judge the significance of quality score differences obtained from the comparison results, as will be shown in Section 2.4.

2.3 Outlier detection

The reliability of the ratings given by a subject can be inspected by checking the transitivity satisfaction rate in his/her comparison results. The transitivity rule is violated when a circular triad is formed among three stimulus, i.e. stimulus i is preferred to stimulus j, stimulus j to k, and stimulus k to stimulus i. A subject whose ratings contain a relatively large number of circular triads is considered as an outlier, and his/her ratings are discarded. This idea was used previously for the binary rating scale without a tie [4], and its extension to accommodate ties is presented here.

Let $i > j$ imply the preference of stimulus i over stimulus j by a subject. And, let $i = j$ mean a tie between the two stimuli. The following four cases are considered as circular triads:

$$\begin{aligned} &i > j \ \cap \ j > k \ \cap \ k > i \\ &i > j \ \cap \ j > k \ \cap \ k = i \\ &i > j \ \cap \ j = k \ \cap \ k > i \\ &i = j \ \cap \ j > k \ \cap \ k > i \end{aligned}$$

Then, the transitivity satisfaction rate is defined by the number of non-circular triads among all possible triads. If its value is less than a threshold for a subject, he/she is considered as an outlier.

In single or double stimulus methodologies, reliability of a subject's ratings is verified by comparing them with those given by other subjects, i.e. if a subject's ratings are deviated too much from the other subjects' ratings, the subject is regarded as an outlier [18]. However, outlier detection in paired comparison can be performed independently without necessity of other subjective data. Moreover, an unreliable subject can be already identified during the test of the subject if his/her ratings already form a considerable amount of circular triads. These features can be particularly beneficial for applying paired comparison to on-line crowdsourced subjective tests.

2.4 Data processing

The ratings of M subjects for intra-group comparison of N stimuli can be summarized by two variables: w_{ij} representing the winning frequency of stimulus i against stimulus j and $t_{ij} = t_{ji}$ indicating the tie frequency between the two stimuli. Note that $w_{ij} + w_{ji} + t_{ij} = M$.

First, let us consider the simplest case where ties are not allowed in the rating scale so that $t_{ij} = t_{ji} = 0$. One of the most popular methods converting the winning frequencies to continuous-scale quality scores is to use the Bradley-Terry (BT) model [3]. In this model, the empirical probability of choosing stimulus i against stimulus j, $P_{ij} = w_{ij}/M$ is represented as

$$P_{ij} = \frac{\pi_i}{\pi_i + \pi_j} \tag{1}$$

where π_i satisfying $\pi_i \geq 0$ and $\sum_{i=1}^{N} \pi_i = 1$ can be considered as the quality score for stimulus i. The parameter π_i can be estimated by maximizing the log-likelihood function given by

$$L(\pi_1, ..., \pi_N) = \sum_{i=1}^{N} \sum_{j=1}^{N} P_{ij} \log \left(\frac{\pi_i}{\pi_i + \pi_j} \right) \tag{2}$$

where $P_{ii} = 0$ by definition.

Modification of the above formulation is required when ties are allowed in paired comparison. A simple way is to treat a tie as half way between the two preference options [7, 15], i.e.

$$w_{ij} \leftarrow w_{ij} + t_{ij}/2 \tag{3}$$

$$w_{ji} \leftarrow w_{ji} + t_{ij}/2 \tag{4}$$

and then the BT model is used as described above.

An explicit extension of the BT model was presented in [19], where an additional parameter $\theta > 1$ is introduced:

$$P_{ij} = \frac{\pi_i}{\pi_i + \theta \pi_j} \tag{5}$$

$$P_{ji} = \frac{\pi_j}{\theta \pi_i + \pi_j} \tag{6}$$

$$P_{i=j} = \frac{\pi_i \pi_j (\theta^2 - 1)}{(\pi_i + \theta \pi_j)(\theta \pi_i + \pi_j)} \tag{7}$$

where $P_{i=j} = t_{ij}/M$ is the empirical probability of ties between stimulus i and j. If $\theta = 1$, the model shrinks to the original BT model. A log-likelihood function similar to (2) can be used to estimate π_i and θ.

Another extension of the BT model is found in [5], which assumes that the probability of no preference is proportional to the geometric mean of the probabilities of preferences:

$$P_{ij} = \frac{\pi_i}{\pi_i + \pi_j + \nu \sqrt{\pi_i \pi_j}} \tag{8}$$

$$P_{ji} = \frac{\pi_j}{\pi_i + \pi_j + \nu \sqrt{\pi_i \pi_j}} \tag{9}$$

$$P_{i=j} = \frac{\nu \sqrt{\pi_i \pi_j}}{\pi_i + \pi_j + \nu \sqrt{\pi_i \pi_j}} \tag{10}$$

where ν is an additional parameter and $\nu = 0$ yields the BT model. The maximum likelihood estimation is used to obtain the parameters π_i and ν.

A tie is provided by a subject when the quality difference of given two stimuli is within the limit of the subject for perceptible quality difference. If more subjects answer ties for the two stimuli, the preferences given by the rest of the subjects can be considered as less significant. The aforementioned methods reflect the quality ambiguity residing in ties to some extent through the scores and additional parameters that are changed according to the tie frequencies. For the method spreading ties to the other preferences and the one in (5)–(7), the scores become similar for a

large number of ties. The parameters θ and ν increase along with an increase of the number of ties. However, such changes do not provide intuitive indexes measuring the ambiguity between the quality of stimuli.

By exploiting the information conveyed in ties about ambiguity or uncertainty of quality difference, we propose a new method that can provide an intuitive measure of confidence information. We assume that quality scores computed from the two preferences (i.e. 'better' and 'worse') may vary within confidence intervals related to the tie probabilities. The upper and lower bounds of the confidence interval are obtained by considering the extreme cases where the ties between two stimuli supposedly belong to one of the two preference options. Therefore, significance of quality difference can be easily verified by examining the amount of overlap between confidence intervals.

The lower and upper bounds of π_i are defined as

$$\pi_i^- = \pi_i - \Delta\pi_i^- \tag{11}$$

$$\pi_i^+ = \pi_i + \Delta\pi_i^+ \tag{12}$$

Then, the interval given by $[\pi_i^-, \pi_i^+]$ becomes the confidence interval around the nominal quality score π_i.

First, π_i is obtained by using the BT model without considering ties, as described above. Then, the imaginary lower and upper empirical probabilities are computed by assuming the ties as either of the two preferences:

$$P_{ij}^- = \frac{w_{ij}}{M} \tag{13}$$

$$P_{ij}^+ = \frac{w_{ij} + t_{ij}}{M} \tag{14}$$

The following equations relate these probabilities and the quality score parameters:

$$P_{ij}^- = \frac{\pi_i^-}{\pi_i^- + \pi_j^+} = \frac{\pi_i - \Delta\pi_i^-}{(\pi_i - \Delta\pi_i^-) + (\pi_j + \Delta\pi_j^+)} \tag{15}$$

$$P_{ij}^+ = \frac{\pi_i^+}{\pi_i^+ + \pi_j^-} = \frac{\pi_i + \Delta\pi_i^+}{(\pi_i + \Delta\pi_i^+) + (\pi_j - \Delta\pi_j^-)} \tag{16}$$

Finally, the additional parameters $\Delta\pi_i^-$ and $\Delta\pi_i^+$ are estimated by solving the maximum likelihood estimation problem for the log-likelihood function given by

$$L'(\Delta\pi_1^-, ..., \Delta\pi_N^-, \Delta\pi_1^+, ..., \Delta\pi_N^+) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left[P_{ij}^- \log\left(\frac{\pi_i - \Delta\pi_i^-}{\pi_i - \Delta\pi_i^- + \pi_j + \Delta\pi_j^+}\right) + P_{ij}^+ \log\left(\frac{\pi_i + \Delta\pi_i^+}{\pi_i + \Delta\pi_i^+ \pi_j - \Delta\pi_j^-}\right)\right] \tag{17}$$

2.5 Global score alignment

From the intra-group comparison, we obtain relative quality scores of the stimuli for the K conditions and their associated lower and upper bounds independently for each group. From the inter-group comparison, relative quality scores of the stimuli of

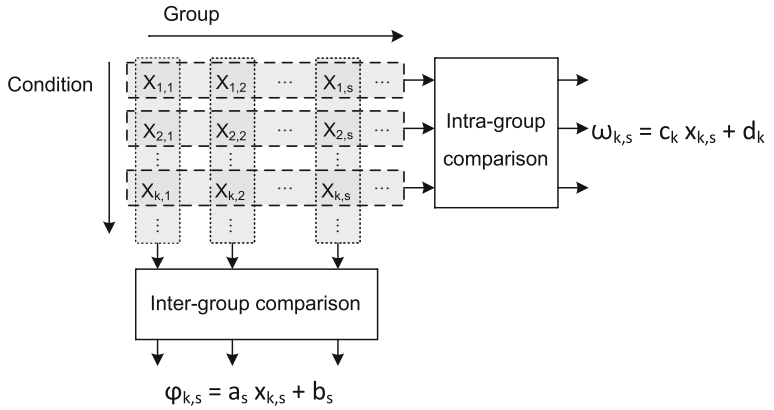

Fig. 1 Process of obtaining intra- and inter-group comparison results from "true" quality scores, which is assumed for global score alignment

the S groups for each of the C ($\leq K$) selected conditions are obtained, e.g. by using the BT model after application of (3)–(4).

Let $\varphi_{k,s}$ ($k = 1, \ldots, K$, $s = 1, \ldots, S$) be the obtained quality score for group s and condition k, which comes from the intra-group comparison. And, let $\omega_{k,s}$ ($k = 1, \ldots, C, s = 1, \ldots, S$) be the quality score obtained from the inter-group comparison for group s and condition k. It can be considered that each of $\varphi_{k,s}$ and $\omega_{k,s}$ is independently drawn from the "true" quality score $x_{k,s}$ that we want to estimate through score alignment. Here, we assume that a linear transform is involved in this drawing process (Fig. 1):

$$\varphi_{k,s} = a_s x_{k,s} + b_s, \quad k = 1, \ldots, K, \ s = 1, \ldots, S \tag{18}$$

$$\omega_{k,s} = c_k x_{k,s} + d_k, \quad k = 1, \ldots, C, \ s = 1, \ldots, S \tag{19}$$

which needs to be solved to obtain aligned quality scores $x_{k,s}$. A constrained optimization technique can be used to solve this system and obtain the globally aligned quality scores $x_{k,s}$.

Since we are interested only in relative scores of stimuli, the maximum and minimum scores after alignment can be set to 100 and 0, respectively. Thus, the total number of equations is $KS + CS$, whereas the number of unknown variables ($x_{k,s}$, a_s, b_s, c_k and d_k) is $KS + 2S + 2C - 2$. In order to solve the problem of global alignment, we need $KS + CS \geq KS + 2S + 2C - 2$, i.e.

$$CS - 2S - 2C + 2 \geq 0 \tag{20}$$

3 Subjective quality assessment of stereoscopic images

This section presents our study on quality assessment of stereoscopic images as an application of the proposed framework. Effects of varying camera distances during

acquisition are discussed based on the obtained results. Especially, we compare them with the results of the single stimulus experiment presented in [8] in order to examine consistency of the results by the two test methodologies, quality discriminability between stimuli, and time complexity.

3.1 Problem definition

As already mentioned in the introduction, performing subjective quality assessment for 3D content is not as straightforward as for 2D content. Many people do not yet have much experience in watching 3D content, and thus do not have general agreement on the definition of perceived quality attributes or cannot easily distinguish between different quality levels. This may result in large variances in the quality ratings of multiple subjects or high ambiguity between stimuli involved in the evaluation, especially when the quality difference between stimuli is small. Moreover, 3D quality is often influenced by a variety of factors interacting in a complicated way. This makes it difficult to evaluate the effect of a specific type of artifacts and its impact on the overall quality.

Many factors are involved in affecting quality of 3D visual content during its acquisition, processing, transmission, and restitution. In particular, acquisition parameters already influence the quality significantly. One of the important acquisition parameters is the distance between the two cameras used for acquiring stereoscopic images. It controls the strength of the 3D effect in a way that a larger distance produces stronger 3D depth impression in the acquired images [20]. However, too extreme depth may cause unnaturalness and even eye discomfort. Therefore, a limited range of the camera distance is usually recommended [12]. A larger distance is allowed for a scene containing a more distant object in order to enhance the depth perception of the object. When the distances of the cameras to the nearest and farthest objects in a scene are known, the maximum permissible camera distance can be theoretically computed [1]. However, many other content features affect the perceived quality of stereoscopic images acquired using different camera distances, e.g. the size and location of the nearest object in the scene and the target viewing condition including the viewing distance and the comfortable viewing range of the considered 3D display [24]. In order to investigate such factors and understand 3D perception of human observers, subjective quality assessment is essential.

3.2 Dataset

The stereoscopic images from the 3D Image Database [8] were chosen for our study. The content varies widely and includes indoor and outdoor scenes with a large variety of colors, textures, and depth structures. Two identical high definition (HD) video cameras mounted on an adjustable stereo mount were used for the image acquisition. Each scene was captured at a resolution of 1920×1080 pixels with 6 different camera distances reaching from 10 cm to 60 cm with a step size of 10 cm (i.e. $K = 6$). Among the ten scenes in the database, 6 scenes were selected for the tests (i.e. $S = 6$), namely, sofa, tables, sculpture, moped, bikes, and construction (Fig. 2). Therefore, we had $6 \times 6 = 36$ stereoscopic images to be evaluated in total. The other scenes were used for training of subjects.

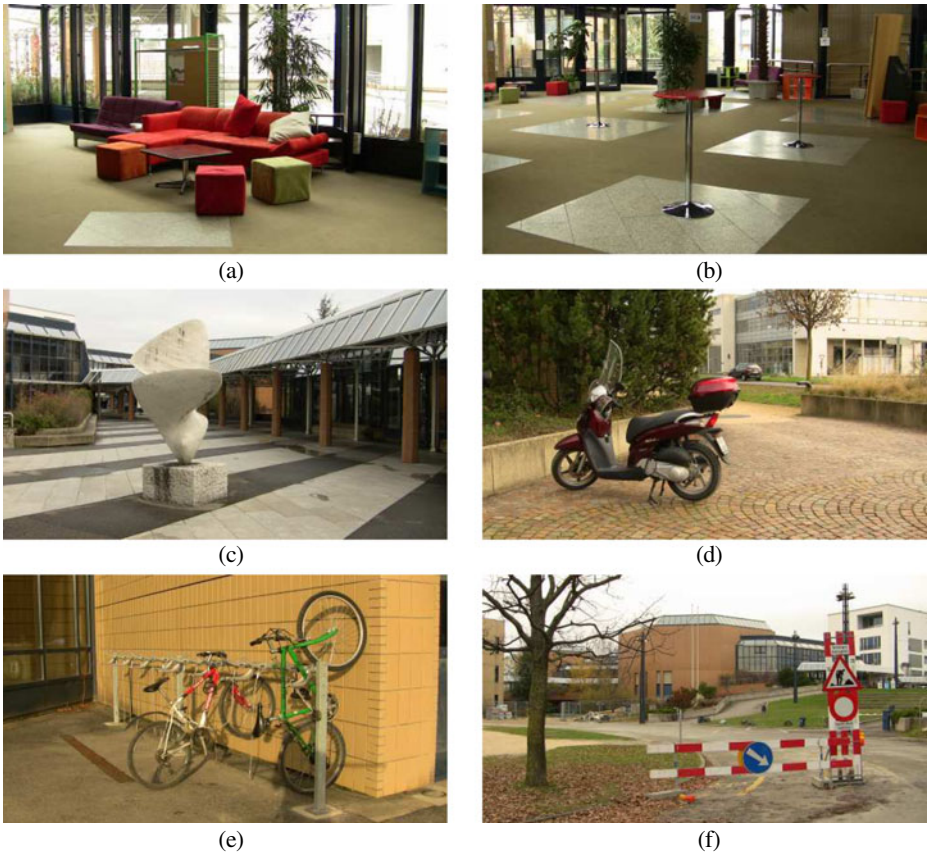

Fig. 2 Example frames of scenes used for the tests: (**a**) sofa, (**b**) tables, (**c**) sculpture, (**d**) moped, (**e**) bikes, and (**f**) construction

3.3 Environment

The subjective tests were conducted in our laboratory designed for professional subjective quality tests according to the recommendations [18]. The ambient lighting consisted of neon lamps of 6,500 K color temperature and the wall color was gray 128.

Two identical 46" LCD polarized stereoscopic monitors with a native resolution of 1920×1080 pixels were used to present two stimuli simultaneously (one on each display). The configuration of the screens and a subject is shown in Fig. 3. Each subject sat in front of the two screens at a distance about 2 m, which is equivalent to approximately 3 times the height of the screens. The subject was allowed to turn his/her head freely to watch the individual stimulus on each screen alternatively. This simultaneous viewing of two stimuli makes the evaluation easier and improves the reliability of the comparison.

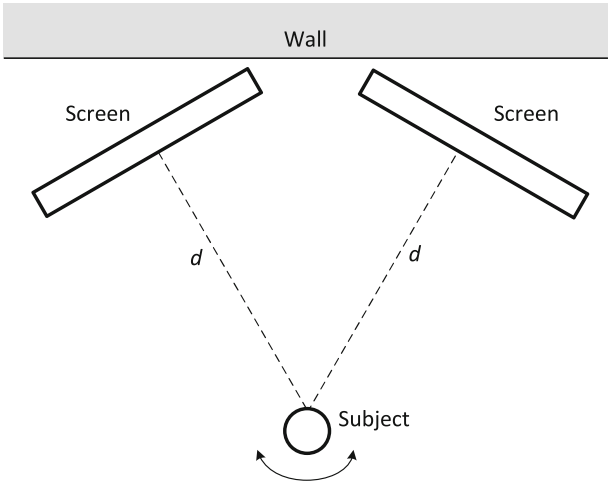

Fig. 3 Configuration of two stereoscopic screens for paired comparison tests. The value of d is approximately 2 m

3.4 Single stimulus test

In [8], the subjective test results based on a single stimulus methodology were reported. Seventeen subjects screened for visual acuity, color vision, and binocular vision participated in the experiment. Each subject sat in front of a 46" LCD polarized stereoscopic screen that is one of the two screens used in our experiment. The distance to the screen was also approximately 2 m. The subject watched each stimulus in a randomized order and then marked its score on an answer sheet. A continuous rating scale ranging from 0 to 100 was used, where the adjective description of each range of the scale ('excellent', 'good', 'fair', 'poor', and 'bad') was also indicated. Outlier detection was performed by following the guidelines given in [18], from which no subject was found as an outlier. From the scores of the seventeen subjects, the MOS and 95% confidence interval was computed for each stimulus.

3.5 Paired comparison test

Sixteen subjects (12 males and 4 females) participated in the tests. They were screened for visual acuity, color vision, and binocular vision [22]. All of them were non-expert viewers with a marginal experience of 3D image and video viewing. Their ages ranged from 25 to 36 with an average of 30.

Prior to the test of a subject, a training session was held to introduce the test methodology to the subject by using a set of training stimuli that were different from the test stimuli. In the middle of the test, a short break was given to prevent the fatigue of the subject.

The outlier detection method described in Section 2.3 was used to detect outliers among the sixteen subjects. Their transitivity satisfaction rates were all above 0.95. Therefore, no subject was rejected as an outlier.

Since the total number of the test stimuli is 36, the number of the full combinations for paired comparison becomes $\binom{36}{2} = 630$, which is infeasibly large for subjective

evaluation. Therefore, the strategy described in the previous section was used to reduce the number of comparisons and still enable estimation of the quality scores of all the test stimuli. One part of the test, i.e. the intra-group comparison, included comparison between different images obtained using different camera distances for the same content, which comprises $6 \times \binom{6}{2} = 90$ image pairs. From the subjective evaluation results of this part, the quality scores of the 6 images were obtained for each scene independently, which was explained in Section 2.4. The other part, i.e. the inter-group comparison, consisted of comparison of 6 different scenes for fixed camera distance parameters. For this, we chose camera distances of 10 cm, 30 cm, and 60 cm, i.e. $C = 3$, which are sufficient to globally align the quality scores for different scenes because (20) is satisfied. Thus, this inter-group comparison includes $3 \times \binom{6}{2} = 45$ pairs. In total, 135 image pairs were evaluated in the test session of each subject. The presentation order of these pairs was randomized and the same content was never shown consecutively.

3.6 Results

3.6.1 Overall results

Figure 4 compares the results of paired comparison with those of the single stimulus experiment after normalization of the scores between 0 and 100. Overall, the results of the two test methodologies show a similar trend, i.e. the best quality for each scene is obtained when the camera distance is small due to comfortable disparity levels, and the quality decreases as the distance increases. For the case of paired comparison, the confidence intervals appear asymmetric around the scores because two separate variables were defined for the lower and upper bounds. This enables us to easily examine the significance of quality difference between two stimuli by checking if the two confidence intervals overlap each other or not. Both methodologies show also the same trend across different scenes. While for some scenes (sofa, moped, bikes, and construction) the quality drops rapidly with increasing camera distance and covers the whole range from 0 to 100, the quality decreases more gradually and does not reach the low quality scores for other scenes (tables and sculpture).

3.6.2 Quality discriminability

An important advantage of paired comparison over the single stimulus methodology is enhanced discriminability between similar quality levels. In [8], it was mentioned that subjects participating in the single stimulus test had difficulty in discriminating quality differences especially for mid-range camera distances. However, such difficulty could be alleviated by simplifying the subjects' task through paired comparison.

In order to compare the two methodologies in terms of discriminability, pairwise comparison results were simulated from the quality scores of the single stimulus experiment as follows. The scores of a subject for each pair of stimuli were compared and the one having a higher score was considered as the winner of the comparison. This was repeated for all subjects and stimuli involved in the single stimulus experiment, from which the empirical preference probabilities P_{ij} were computed. Figure 5 shows the preference probabilities from the single stimulus and the paired comparison tests. Discriminability between quality of stimuli can be

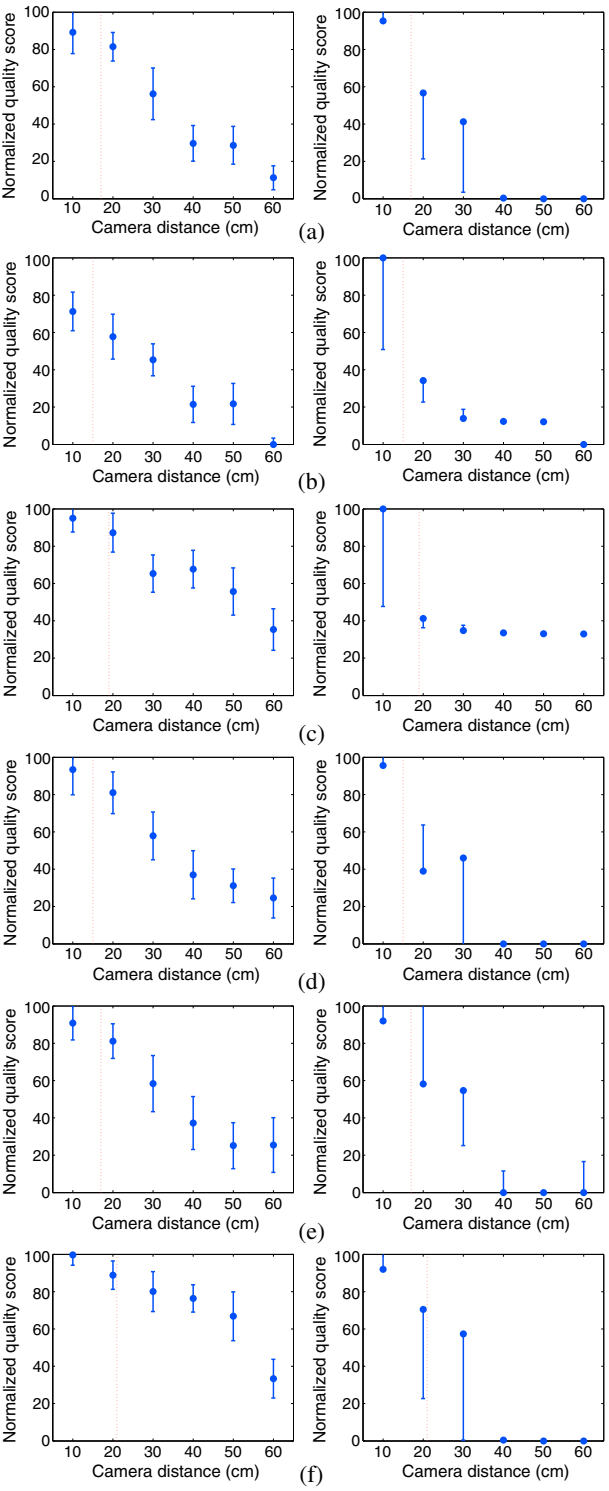

Fig. 4 Normalized quality scores obtained from single stimulus (*left column*) and paired comparison (*right column*) experiments. The theoretical camera distance limits are shown as *dotted red lines*. (**a**) sofa (**b**) tables (**c**) sculpture (**d**) moped (**e**) bikes (**f**) construction

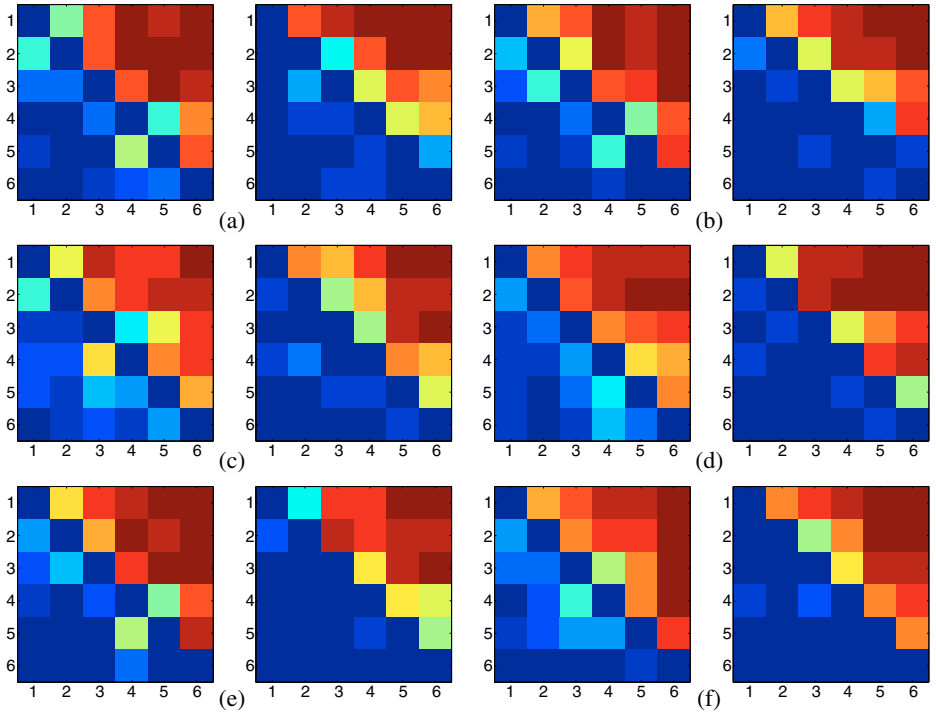

Fig. 5 Preference probability choosing stimulus i (x-axis) against stimulus j (y-axis) for the single stimulus (*left column*) and paired comparison (*right column*) experiments. *Red* and *blue colors* indicate high and low probabilities, respectively. (**a**) sofa (**b**) tables (**c**) sculpture (**d**) moped (**e**) bikes (**f**) construction

inferred by comparing P_{ij} and P_{ji} (symmetric with respect to the main diagonal), i.e. if they have similar values, their quality levels are more ambiguous. It is clearly observed that the paired comparison results show smaller ambiguity than the single stimulus results, especially for the neighboring pairs having similar camera distances and consequently similar quality levels. Especially, for medium camera distances there is a lot of ambiguity between adjacent camera distance values in the case of the single stimulus methodology, which is largely reduced in the paired comparison methodology.

In simulating paired comparison from the single stimulus results, ties can be generated by treating an absolute score difference smaller than a threshold as a tie. Then, a measure of quality discriminability for stimulus i and stimulus j is defined as

$$D_{ij} = |P_{ij} - P_{ji}| \tag{21}$$

The overall discriminability measure D can then be computed as the average across all stimulus pairs. By adjusting the threshold producing the ties, the simulated results having the same tie probabilities to those of paired comparison were obtained from the single stimulus results. Figure 6 compares the average discriminability measures of the two methodologies for each scene. It can be seen that, except for sofa, the

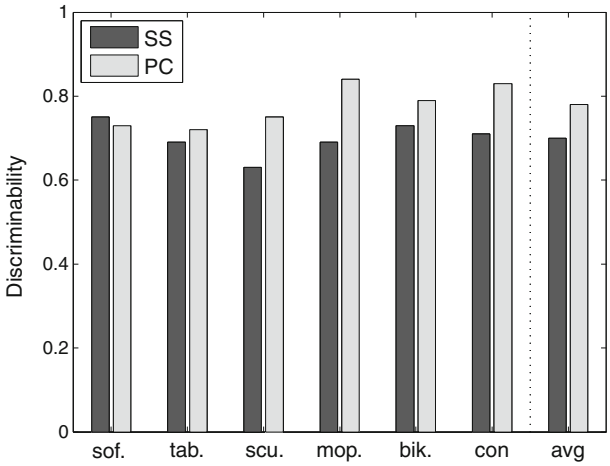

Fig. 6 Discriminability measures of the single stimulus (SS) and paired comparison (PC) methodologies for the same tie probabilities

paired comparison always outperforms the single stimulus methodology in terms of the discriminability.

Our finding about the effectiveness of paired comparison in distinguishing different 3D quality levels is also in line with the observation presented in [2]: The absolute categorical rating and paired comparison methods were compared for quality assessment of view synthesis algorithms in terms of the required minimum number of subjects that allow the statistical distinction between stimuli, from which it was shown that quality difference between stimuli could be statistically identified with less subjects through paired comparison than absolute categorical rating.

3.6.3 Complexity

The improved discriminability between quality levels in paired comparison is obtained at the expense of an increased time complexity. A single stimulus test included the evaluation of 36 stimuli, whereas 135 stimulus pairs were evaluated in a paired comparison test. Each stimulus (or stimulus pair) was shown for 8 s and then 4 s were given for its rating. Additionally, five dummy stimuli were presented at the beginning of each test for stabilization of the subjective evaluation. Therefore, a single stimulus test took $(5 + 36) \times (8 + 4) = 492$ s, while a paired comparison test took $(5 + 135) \times (8 + 4) = 1680$ s. Nevertheless, post-test interviews of the subjects revealed that paired comparison significantly relieved the difficulty of rating on a continuous scale because of the simplified rating scale and simultaneous viewing of two stimuli. In addition, the training procedure was much simpler and shorter in time for the paired comparison experiment. Considering such advantages, the increased time complexity of paired comparison can be justified.

4 Conclusion

We have presented a new analysis method for paired comparison-based subjective quality assessment, which was applied to quality evaluation of stereoscopic images. The complete quality evaluation methodology including the test procedure, outlier

detection, score calculation, and significance analysis was described. Exploiting the ambiguity residing in ties, the proposed method facilitates intuitive examination of significant quality difference without any additional statistical hypothesis test. The method was used for the subjective evaluation of stereoscopic images acquired with varying camera distances. It was demonstrated that, while the results are similar for both single stimulus and paired comparison tests, the paired comparison test methodology improves quality discriminability between stimuli.

As discussed, the advantages of paired comparison are obtained at the cost of increased time complexity in comparison to other methods such as the single and double stimulus test methodologies. Thus, improving the efficiency of paired comparison by reducing its time complexity is an important further research topic in order to widen its applicability to various quality assessment problems. There exist some recent work on this, e.g. [6]. Our future work will explore extension of the proposed analysis method for complexity-decreased paired comparison tests. We also plan to apply the proposed method to other evaluation tasks where conventional single or double stimulus methodologies have difficulty in obtaining reliable results.

Acknowledgements This work was supported in part by the Ministry of Knowledge Economy, Korea, under the IT Consilience Creative Program (NIPA-2010-C1515-1001-0001), in part by Yonsei University Research Fund of 2011, and in part by the COST Action IC1003 European Network on Quality of Experience in Multimedia Systems and Services (Qualinet).

References

1. Bercovitz J (1998) Image-side perspective and stereoscopy. In: Proc. SPIE, vol 3295
2. Bosc E, Pépion R, Callet PL, Köppel M, Ndjiki-Nya P, Pressigout M, Morin L (2011) Towards a new quality metric for 3-D synthesized view assessment. IEEE J Sel Topics Signal Process 5(7):1332–1343
3. Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika 39:324–345
4. Chen KT, Wu CC, Chang YC, Lei CL (2009) A crowdsourceable QoE evaluation framework for multimedia content. In: Proc. ACM multimedia, pp 491–500
5. Davidson RR (1970) On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. J Am Stat Assoc 65(329):317–328
6. Eichhorn A, Ni P, Eg R (2010) Randomized pair comparison- an economic and robust method for audiovisual quality assessment. In: Proc. int. workshop on network and operating systems support for digital audio and video. Amsterdam, The Netherlands, pp 63–68
7. Glickman ME (1999) Parameter estimation in large dynamic paired comparison experiments. J R Stat Soc, Ser C, Appl Stat 48(3):377–394
8. Goldmann L, Simone FD, Ebrahimi T (2010) Impact of acquisition distortions on the quality of stereoscopic images. In: Proc. int. workshop on video processing and quality metrics for consumer electronics. Scottsdale, Arizona, USA
9. Gotchev A, Akar GB, Capin T, Strohmeier D, Boev A (2011) Three-dimensional media for mobile devices. Proc IEEE 99(4):708–741
10. Huynh-Thu Q, Barkowsky M, Callet PL (2011) The importance of visual attention in improving the 3D-TV viewing experience: overview and new perspectives. IEEE Trans Broadcast 57(2):421–431
11. Huynh-Thu Q, Callet PL, Barkowsky M (2010) Video quality assessment: from 2D to 3D-challenges and future trends. In: Proc. int. conf. image processing. Hong Kong, China, pp 4025–4028
12. Ijsselsteijn WA, de Ridder H, Vliegen J (2000) Subjective evaluation of stereoscopic images: effects of camera parameters and display duration. IEEE Trans Circuits Syst Video Technol 10(2):225–233

13. Jumisko-Pyykkö S, Utriainen T (2011) A hybrid method for quality evaluation in the context of use for mobile (3D) television. Multimed Tools Appl 55(2):185–225
14. Kooi FL, Toet A (2004) Visual comfort of binocular and 3D displays. Displays 25(2–3):99–108
15. Lee JS, Simone FD, Ramzan N, Zhao Z, Kurutepe E, Sikora T, Ostermann J, Izquierdo E, Ebrahimi T (2010) Subjective evaluation of scalable video coding for content distribution. In: Proc. ACM multimedia. Firenze, Italy, pp 65–72
16. Liu X, Yang LT, Sohn K (2011) High-speed inter-view frame mode decision procedure for multi-view video coding. Future Gener Comput Syst. doi:10.1016/j.future.2011.05.013
17. Meesters LMJ, Ijsselsteijn WA, Seuntiëns PJH (2004) A survey of perceptual evaluations and requirements of three-dimensional TV. IEEE Trans Circuits Syst Video Technol 14(3):381–391
18. International Telecommunication Union (2002) Methodology for the subjective assessment of the quality of television pictures. Recommendation ITU-R BT.500-11
19. Rao PV, Kupper LL (1967) Ties in paired-comparison experiments: a generalization of the Bradley-Terry model. J Am Stat Assoc 62(317):194–204
20. Seuntiiëns P, Meesters L, Ijsselsteijn W (2006) Perceived quality of compressed stereoscopic images: effects of symmetric and asymmetric JPEG coding and camera separation. ACM Trans Appl Percept 3(2):95–109
21. Smolic A, Kauff P, Knorr S, Hornung A, Kunter M, Müller M, Lang M (2011) Three-dimensional video postproduction and processing. Proc IEEE 99(4):607–625
22. International Telecommunication Union (2000) Subjective assessment of stereoscopic television pictures. Recommendation ITU-R BT.1438
23. International Telecommunication Union (1999) Subjective video quality assessment methods for multimedia applications. Recommendation ITU-R P.910
24. Zilly F, Kluger J, Kauff P (2011) Production rules for stereo acquisition. Proc IEEE 99(4):590–606

Jong-Seok Lee received his Ph.D. degree in electrical engineering and computer science in 2006 from KAIST, Korea, where he also worked as a postdoctoral researcher and an adjunct professor. From 2008 to 2011, he worked as a research scientist in the Multimedia Signal Processing Group (MMSPG) at Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland. Now he is an assistant professor at the School of Integrated Technology in Yonsei University, Korea. His current research interests include audio-visual signal processing, multimedia quality assessment and multimodal human-computer interaction. He has been or is involved in several national and European projects as well as industry projects. He has been also involved in international standardization activities in MPEG and JPEG. He was the chair of the First Spring School on Social Media Retrieval (S3MR) held in Interlaken, Switzerland, in 2010, and an organizing committee member of the Second Summer School on Social Media Retrieval (S3MR) held in Antalya, Turkey, in 2011. He is author or co-author of over 40 publications. He is a voting member of the Multimedia Communication Technical Committee (MMTC) of the IEEE Communication Society.

Lutz Goldmann received his Dipl.-Ing. (M.Sc.) degree in electrical engineering from the Technical University of Dresden, Germany in 2002 and Dr.-Ing. (Ph.D) degree in electrical engineering from the Technical University of Berlin (TUB), Germany in 2009. In 2002, he joined Siemens CT IC2, Munich, Germany, as a research student, where he developed image enhancement techniques for video coding artifact removal. Between 2003 and 2008, he worked as a research assistant at TUB on the detection and recognition of humans within images and videos. He has been actively involved in several national and European projects, such as GrVis, VISNET, 3DTV, K-Space, and VISNET II. He is now a research scientist at Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. He is author or co-author of more than 20 research papers. His research interests include 2D and 3D image and video analysis, multimedia quality assessment, and machine learning.

Touradj Ebrahimi received his M.Sc. and Ph.D., both in electrical engineering, from Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1989 and 1992, respectively. From 1989 to 1992, he was a research assistant at the Signal Processing Laboratory of EPFL. During the summer 1990, he was a visiting researcher at the Signal and Image Processing Institute of the University of Southern California, Los Angeles, California. In 1993, he was a research engineer at the Corporate Research Laboratories of Sony Corporation in Tokyo. In 1994, he served as a research consultant at AT&T Bell Laboratories. He is now a professor at EPFL heading the Multimedia Signal Processing Group (MMSPG). He is also an adjunct professor with the Center of Quantifiable Quality of Service at Norwegian University of Science and Technology (NTNU), Trondheim, Norway. He is involved with various aspects of digital video and multimedia applications. He is author or co-author of more than 200 papers and holds 14 patents.